词元经济

AI时代财富增长密码

蒋凡 著

文化发展出版社
Cultural Development Press

词元经济

AI时代财富增长密码

蒋凡 著

文化发展出版社
Cultural Development Press

版权信息

书名：词元经济：AI时代财富增长密码

作者：蒋凡

出版社：文化发展出版社·果麦文化

出版时间：2026年6月

ISBN：9787514250770

字数：86千字

本书由果麦文化传媒股份有限公司授权得到APP电子版制作与发行

前言

不做词元消费者，要做词元生产者

2026年3月16日，阿里巴巴专门成立Token事业群，由集团CEO吴泳铭直接负责，旨在以“创造Token、输送Token、应用Token”为核心目标，全面整合集团AI资源。

3月17日，英伟达CEO黄仁勋在GTC年度大会主题演讲中首次提出“Token Economy”，向业界描绘了如何深入理解Token训练推理机制，通过工业级优化及调用Token，运用智能生产力创造价值的新商业模式。在他的预判中，未来的数据中心将转变为全天候运转的Token工厂，以数据与电力作为核心原材料，以Token作为最终产出。而Token这一产品，最终会像石油、电力一样，成为支撑整个世界运转的基础能源。

3月23日，国家数据局正式做出回应，Token被确认译为“词元”，并给出了官方定义：可计量、可定价、可交易的标准化商品。这一决定不仅终结了过去数年中该领域的术语混乱状态，更让这一AI时代的核心概念正式走进了大众的视野，为整个产业的标准化发展奠定了基础。

这几个事件，共同标志着词元经济时代已经正式到来。然而绝大多数的普通人，乃至部分行业从业者，却尚未充分意识到，这一全新的时代究竟意味着怎样的产业变革与财富机遇。

如今打开任何一个社交平台或是资讯网站，映入眼帘的信息几乎都在围绕着如何使用AI工具展开：如何利用ChatGPT撰写文案，如何借助Midjourney完成设计工作，如何通过AI接单赚钱，都在向人们传递着一个核心信息：在AI时代，必须学会使用AI工具，否则就会被时代淘汰。然而，这里隐藏着对AI时代本质的认知偏差。

这些信息想要引导人们完成的，仅仅是成为词元的消费者，成为AI时代的从业者。人们一股脑儿地参与进去，在使用AI工具的过程中消耗词元，进而通过出卖自身的时间获取有限的服务收入，往往就会认为自己已经抓住了AI的产业红利，但实际上却仅仅是在为上游的词元生产者提

供劳动力，成为词元经济的下游打工人，甚至是语料喂养者。

让我们假设一下，如果你使用ChatGPT撰写一份商业文案，需要向平台支付一元的词元费用，随后向客户收取十元的服务费用，你会认为自己赚取了九元的利润。但实际上，ChatGPT平台生产这些词元的成本只有一角钱，平台赚取了九角的毛利，你花费了一个小时的时间，仅仅获得了九元的收入。在这个过程中，你只是付出了时间成本的打工人，而平台才是掌握了生产能力的经营者。

这就是当前AI行业的真相：绝大多数人都在下游环节成为词元的消费者，成为AI时代的普通从业者，而只有少数人在上游环节掌握生产资料，成为词元的生产者，成为新的智产阶级。他们掌握了词元的生产能力，掌握了产业的定价权，他们才是这个新时代的真正赢家，才是这个时代财富的真正拥有者。

回顾人类历史上的历次技术革命，我们能够看到这一相同的现象反复发生：

在农耕时代，绝大多数人都是从事农业生产的农民，而只有少数人拥有土地资源，他们成为地主，掌握了整个社会的核心财富。

进入工业时代，绝大多数人都是工厂的工人，依靠出卖劳动力获取收入，而只有少数人拥有工厂与机器设备，他们成为资本家，掌握了整个时代的核心财富。

到了互联网时代，绝大多数人都是平台的用户，成为流量的消费者，而只有少数人拥有互联网平台，掌握了流量的分发权，他们成为互联网行业的巨头，掌握了整个时代的核心财富。

而在当前的AI时代、词元经济时代，这一规律依然没有改变。绝大多数人最终都会成为词元的消费者，成为AI时代的普通从业者；而只有少数人会成为词元的生产者，拥有属于自己的词元生产线，他们会掌握这个时代的核心财富，成为这个时代的全新巨头。

这就是本书希望向读者传递的核心信息，本书不会教读者如何使用AI工具，不会引导读者通过AI工具获取服务收入，不会引导读者成为AI时代的从业者。本书的核心目标，是向读者讲述如何成为词元的生产者，如何搭建属于自己的词元生产线，如何拥有属于自己的印钞机，如何在这

个全新的时代掌握属于自己的财富增长密码。

不少读者或许会产生疑问，词元生产难道不是科技巨头的专属领域吗？难道不是需要投入数十亿的资金，组建数千人的技术团队才能够完成的工作吗？作为普通人，又怎么可能参与到这一领域当中？

事实上，这一认知已经与当前的产业现状脱节。开源模型的成熟落地、推理优化技术的快速进步、全球模型交易市场的完善，已经将搭建词元生产线的门槛降低到了极致。当前，哪怕是一名普通的个人开发者，哪怕只拥有数千元的初始资金，也能够搭建属于自己的词元生产线，也能够将自己生产的词元销售给全球的用户，也能够获取持续稳定的被动收入。

这并非什么天方夜谭式的空想，而是当前正在真实发生的产业实践。这就是词元经济为普通大众带来的全新机遇，它让人们不需要投入大量的初始资金、不需要组建庞大的技术团队、不需要拥有复杂的行业资源、只要掌握了正确的方法，就能够搭建属于自己的生产线，就能够拥有属于自己的印钞机。

本书将从最基础的词元定义、词元经济的基础概念开始，一步一步地带领读者看懂这个全新的时代，带领读者搭建属于自己的词元生产线，带领读者将生产线转化为持续盈利的印钞机，带领读者抓住这个时代的最大机遇。

整本书的内容，共分为以下五个核心部分。

第一部分，看懂词元。我们会向读者阐释词元的核心定义，词元经济的三层产业格局，帮助读者明确自身的产业站位；同时解读词元经济的合规性框架，向读者说明这一赛道是国家明确支持的合法赛道，不存在任何的政策风险，解除后顾之忧。

第二部分，词元经济的铁三角。我们会向读者阐释词元生产的核心逻辑，电力、算力、词元这三个核心要素的运作机制，帮助读者理解如何以最低的成本生产最多的词元。我们会告诉读者词元经济的两种核心财富路径：低成本的规模优势路径与高价值的专业优势路径；同时解读规模与复利的终极真相，帮助读者理解如何实现一次搭建、永久盈利的终极目标。

第三部分，帮助读者搭建第一条词元生产线。我们会带领读者一步步地完成生产线的搭建，从硬件配置到算法优化，从模型选择到交付体系，从低成本的通用生产线到高价值的专业生产线，同时介绍零代码的搭建方法，确保哪怕是没有技术基础的读者也能够完成生产线的搭建。

第四部分，词元商业化。我们会向读者阐释如何将生产线转化为“印钞机”、上游的五大主流商业模式、低成本的规模化增长路径、高价值的长期订单利润、词元出海的全球化机遇，以及智能体时代的增量爆发，帮助读者获取最多的订单，赚取最高的利润。

第五部分，未来趋势。我们会站在产业发展的长期视角，向读者阐释词元经济的终极格局、词元如何成为社会的基础能源、全球产业链的重构路径、下一代词元工厂的发展形态、个人与企业的终极财富形态，以及投资与布局的优先级，帮助读者抓住最确定的长期机遇。

这五个部分共同构成了完整的词元经济知识体系。我们会以最通俗的语言、最真实的产业案例、最落地的实践方法，将所有的内容完整地呈现给读者，读者不需要拥有任何的技术基础，只要跟随本书的步骤，就能够学会如何搭建自己的生产线，如何在这个全新的时代获取属于自己的财富。

读者或许会产生疑问，当前阶段入场，是否已经错过了最佳的时间窗口？

我们需要明确的是，当前阶段入场一点都不晚，词元经济的时代才刚刚开始。我们走完的整个产业发展进程还不到10%，未来还有90%的发展空间，还有无数的机遇等待着人们去挖掘。

过去互联网时代的红利，很多人错过了；移动互联网时代的红利，很多人也错过了。但是这一次，词元经济的时代，我们不能再错过了，因为这是人类历史上最大的财富浪潮，它比互联网、比移动互联网的市场规模还要大十倍、百倍，它会彻底地改变整个世界的经济、产业与财富逻辑，它会创造出无数的全新财富与全新机遇。

当前就是最好的入场时机。当前入场，就能够吃到整个产业浪潮的最大红利，就能够在这个全新的时代站稳脚跟，就能够实现属于自己的财富自由。

我撰写这本书的核心目标，就是希望能够帮助读者抓住这个机遇。希望大家不要像过去很多次一样，错过了这个时代的红利；希望大家能够成为词元的生产者，而不是消费者；希望大家能够拥有属于自己的“印钞机”，而不是为别人打工。因为在词元经济的时代，只有距离货币最近的人，只有掌握货币生产流程的人，才能够拥有最大的财富，这是所有的技术革命、所有的时代都不曾改变的真理。

现在翻开这本书，跟着我一起走进词元经济的世界，一起搭建你的第一条词元生产线，一起抓住这个时代的最大机遇，一起解锁AI时代的财富增长密码。

第一部分

看懂词元：
财富的新计算单位

当人工智能的浪潮席卷全球，大模型重构产业逻辑时，绝大多数的市场参与者都陷入了对AI的浅层认知误区：他们将人工智能视为可直接调用的效率工具，将词元视为调用工具时消耗的成本，将AI时代的财富机会等同于依托工具承接外包服务，却从未意识到，词元并非只是模型处理信息的最小语义单元，而且是人工智能时代全新的价值载体，是支撑整个智能社会运转的基础能源，是重构全球财富分配逻辑的核心要素。

这一认知的偏差，让很多人在时代的浪潮中站错了位置，他们在下游的消费环节内卷，却从未意识到，上游的生产环节，才是这个时代真正的财富源头。在这一部分，我们将从最基础的定义出发，带你穿透技术的表象，深入理解词元的本质与价值，拆解词元经济的完整产业格局，帮你找到最适合自己的产业站位，同时，我们也将为你解读词元经济的合规性框架，消除你对政策风险的顾虑，让你能够在安全、合法的赛道上，放心地布局与发展。

这是你进入词元经济领域的第一堂课，我们将帮你打破固有的认知，建立对词元经济的完整、系统的基础认知，只有彻底搞懂了这些基础的逻辑，你才能够在后续的内容中，真正理解词元生产的核心逻辑，理解财富增长的底层路径，理解未来的产业发展趋势，真正抓住这个时代的最大机遇，在人工智能的浪潮中，占据属于自己的位置。

第一章 词元是**AI**时代的货币

2026年3月23日，北京钓鱼台国宾馆，中国发展高层论坛2026年年会的现场座无虚席。国家数据局局长刘烈宏走上讲台，将AI领域核心术语Token的中文标准译名，正式表述定为“词元”，并明确了它的定义和产业意义。

这短短一句话，终结了过去数年里行业内关于“令牌”“代币”“通证”等译名的混乱局面，更重要的是，它将这个原本只存在于AI技术圈层的专业术语，正式推到了大众视野的中央，成为整个智能时代价值体系的新锚点。

“词元不仅是智能时代的价值锚点，更是连接技术供给与商业需求的结算单位，为商业模式的落地提供了可量化的可能”，刘烈宏在演讲中如此定义这一命名的意义。在他看来，词元将不再仅仅是大模型处理语言的一个技术单元，它将成为智能时代的“新货币”，成为衡量AI服务价值、结算商业交易、分配数字财富的核心单位。

这一幕让人不禁联想起历史上每一次经济时代更迭时，出现的那些新的价值符号，它们必将改变亿万人的命运。

1.1 从贝壳到石油：人类经济的价值单位迭代

回望人类经济发展的数千年历史，我们会发现一个清晰的规律：每一个时代，都有属于那个时代的“硬通货”，都有一个被全社会共同认可的，用来衡量财富、进行交易的核心单位。

在农耕文明的早期，人类社会的生产力极度低下，物资的匮乏使得交换行为极为原始。此时，贝壳因为其坚固耐用、易于携带、难以伪造的特性，成为最早的一般等价物。一个贝壳，就是当时社会的最小价值单位，人们用它来衡量粮食的多少、布匹的价值，甚至是奴隶的价格。拥有贝壳的多少，直接决定了一个人在那个时代的财富水平。

进入大航海时代，全球贸易的兴起使得跨区域的商品流动成为常态。此时，白银和黄金取代了贝壳，成为新的硬通货。西班牙殖民者从美洲开采的银圆，沿着贸易航线流遍全球，成为国际贸易的结算货币。一个银

圆，代表着一定重量的白银，它可以在欧洲买到工业品、在亚洲买到香料和丝绸、在美洲买到原材料。谁掌握了银圆的铸造和流通，谁就掌握了那个时代的财富分配权。

而到了工业时代，情况又发生了变化。随着蒸汽机的发明和工厂的普及，能源成为驱动整个社会运转的核心动力。石油，这个被称为“工业的血液”的黑色黄金，成为新的核心价值单位。一桶原油的价格，成为全球经济的晴雨表；一个国家拥有石油的多少，直接决定了其工业的竞争力。从汽车到飞机，从塑料到化纤，几乎所有的工业产品，其成本核算的核心都离不开石油。谁控制了石油的生产和运输，谁就拥有了影响全球经济的话语权。

互联网时代，这一规律再次上演。当信息革命席卷全球，流量成为新的财富密码。一个点击、一个用户、一个下载量，成为互联网公司衡量价值的核心指标。腾讯、阿里、字节跳动这些巨头，无一不是通过掌握海量的流量，构建起了自己的商业帝国。在那个时代，谁拥有了流量，谁就拥有了用户，谁就拥有了变现的基础。流量的多少，直接决定了一家公司的估值，决定了广告的定价，决定了电商的转化。

而今天，当我们站在AI时代的入口，我们看到了历史的重演。一个新的价值单位正在接过接力棒，成为这个新时代的硬通货。它，就是词元。

1.2 词元：**AI**处理信息的最小语义单元

要理解词元为什么能成为AI时代的货币，我们首先要回到技术的本源，搞清楚：到底什么是词元？

从技术层面来说，词元是大型语言模型（LLM）处理文本和代码的最小、不可再分割的语义单元。

很多人误以为，词元就是词语，或者就是汉字。但实际上，词元的划分远比这要复杂。人类的语言在AI模型内部，并非以我们日常看到的字符或者词语的形式被直接处理，而是通过一个名为“分词器”（Tokenizer）的工具，被切分成一系列数值化的“词元”序列。

举个简单的例子，当我们输入英文句子“Hello, world! ”，分词器并不会把它当成一整个句子，也不会简单地按空格拆分，而是会把它切分成 “Hello”“, ”“world”“！ ”四个独立的词元。而对于中文句子“我爱中

国”，分词器可能会把它切分成“我”“爱”“中国”三个词元。

这里有一个很有意思的细节，词元的划分并非简单地按字或词，而是基于模型的算法和词汇表（Vocabulary）进行动态决策的。一个词元，可能对应一个完整的词语，比如“中国”；也可能只是一个词的一部分，比如一些生僻字，可能会被拆分成2—3个词元来处理；甚至，它可能只是一个标点符号，或者一个空格。

这背后的逻辑是什么？其实是为了在信息的压缩率和处理的效率之间找到一个最优的平衡点。如果我们把每个字符都当成一个独立的单元，那么序列会变得很长，处理起来效率很低。但如果我们把整个词语都当成一个单元，那么词汇表的规模就会变得非常庞大，难以管理。而词元的机制，完美地解决了这个问题：它把常用的词语、短语合并成一个词元，把生僻的字符拆分成更小的单元，这样既控制了词汇表的规模，又提升了处理的效率。

正是因为有了词元这个最小单元，大模型才能理解人类的语言，才能进行推理，才能生成内容。无论是用户输入的提示词，还是模型生成的文章、代码，最终都会被切分成一系列词元，然后在模型内部进行计算和处理。可以说，整个大模型的运转，本质上就是对词元的处理和流转。

但词元的意义，绝不仅仅止步于此。如果它只是一个技术单元，那它永远都只会停留在实验室里，不会成为整个时代的价值锚点。词元真正的革命性意义，在于它天生就具备成为货币的三大核心属性：可计量性、可定价性、可交易性。

1.3 可计量、可定价、可交易：词元的货币属性

这三大属性，是词元能够成为AI时代硬通货的核心基石。

首先是可计量性。

在过去，AI服务的价值一直是模糊的。当用户使用AI工具的时候，我们很难说清楚，他到底消耗了多少资源，这个服务到底值多少钱。有的公司按调用次数收费，有的按算力时长收费，有的干脆卖终身授权。这种模糊的计量方式，使得AI服务的成本核算变得异常困难，也严重制约了其商业化的普及。

但词元改变了这一切。任何一项AI服务，无论是文本生成、图像识别还是代码编写，其消耗的计算资源都可以被精确地统计为词元的数量。使用者输入了100个词元，模型输出了200个词元，那么这次服务总共消耗了300个词元。这个数字是精确的、是可量化的、是没有任何模糊空间的。

这就像什么？就像我们家里的电表。过去，电网公司可能不知道用电户用了多少电，只能按月交固定的电费。但有了电表之后，使用者用了一度电，电表就走一度，精确到小数点后两位。词元就是AI时代的“电表”。它把无形的智能服务变成了有形的、可精确度量的数字。

词元的计量主要发生在两个环节：输入（Input）和输出（Output）。当用户向大模型发送请求时，系统会首先将用户的输入（如文本、图片等）通过分词器（Tokenizer）转换为词元。然后，在模型生成响应的过程中，每产生一个词元（无论是字符、词语还是子词），都会被计为输出词元。

绝大多数AI服务商采用双向计费的模式，即输入词元和输出词元都会被计入总消耗，但两者的单价通常不同，输出词元的单价一般高于输入词元，因为生成内容的计算成本通常高于理解内容的成本。计量的精确性至关重要，它不仅影响用户的付费，也决定了模型服务商的收入。因此，各大平台都在其API文档中明确规定了计费的具体方式，并提供了调用次数、输入/输出词元数量等详细的用量统计，以确保计费的透明与公正。

正是因为有了这种可计量性，AI服务的成本核算才变得清晰起来。一个企业可以精确地知道，全公司的员工每天在AI应用上消耗了多少词元，每个词元的成本是多少，整个业务的边际成本是多少。这在过去，是根本无法想象的。

其次是可定价性。

当一个东西可以被精确计量的时候，它就具备了被定价的基础。

词元的出现使得AI服务商可以像水电煤一样，按照用户消耗的词元数量进行透明、标准的计费。使用者用了多少，就付多少钱。用得多，付得多；用得少，付得少。这彻底改变了过去AI服务那种模糊的、打包式的收费模式。

我们现在已经能看到这种模式的普及。OpenAI的GPT模型，输入词元每百万个收费0.15美元，输出词元每百万个收费0.6美元；Anthropic的Claude，价格稍微高一点；而中国的MiniMax，输入价格只有0.3美元每百万词元，不到Claude的1/17。这些价格，都是基于词元来计算的。

用户在使用这些服务的时候，就像交水电费一样简单明了。我今天用了多少词元，大概要花多少钱，一目了然。这种清晰的定价模式，极大地降低了中小企业和个人开发者使用AI的门槛。过去，人们要想用AI，可能要花几十万买一套软件，或者花几百万做一个项目。但现在，人们可以按用量付费，用多少付多少，几块钱也能开始用。这就把AI从少数大企业的玩具，变成了所有人都能用得起的普惠工具。

最后是可交易性。

当一个东西既可以计量，又可以定价的时候，它自然就具备了交易的属性。

词元可以在产业链的各个环节之间自由地流转、分配和结算。上游的算力提供商，把算力卖给模型厂商，模型厂商用算力生产词元；中游的平台服务商把词元打包，卖给下游的应用开发商；下游的应用开发商再把词元封装成服务，卖给最终的用户。

在这个过程中，词元就像货币一样在整个产业链里流通。算力厂商可以用它来结算，模型厂商可以用它来核算成本，应用厂商可以用它来给用户定价。甚至，未来我们可能会看到，词元可以像大宗商品一样，在交易市场上进行买卖，甚至投机。有人生产了多余的词元，可以卖给需要的人；有人需要词元的时候，也可以从市场上购买。

这就形成了一个完整的商业闭环。从算力到词元，从生产到消费，整个产业链都围绕着词元这个核心单位来运转。这就是词元经济的基本形态。

1.4 词元：**AI**时代真正的硬通货

正是因为具备了这三大属性，词元才成为AI时代真正的硬通货。美国英伟达公司创始人兼总裁黄仁勋用一句话点破了这一点：“词元是新的大宗商品。”

大宗商品，这个词我们都很熟悉。石油、天然气、铁矿石、粮食，这些都是大宗商品。它们是整个工业社会的基础原材料，是可以标准化、大规模交易、价格波动会影响整个经济的核心商品。可词元为什么会成为AI时代的大宗商品？

因为它和传统的大宗商品太像了。

首先，它是标准化的。一个词元，不管是谁生产的，不管是用什么模型生产的，它的计量标准是统一的。这就像一桶石油，不管是哪个油田产的，它的计量单位都是桶，它的热值都是标准化的。词元也是一样，国家数据局已经统一了它的定义，统一了它的计量标准，这就为它的大规模交易奠定了基础。

其次，它是刚需。就像工业时代离不开石油一样，AI时代离不开词元。任何一个AI应用，任何一个智能服务，都必须消耗词元才能运转。没有词元，大模型就无法工作，AI就无法提供服务。不管是智能客服，还是内容创作，不管是自动驾驶，还是工业质检，所有的智能场景，本质上都是在消耗词元。

最后，它的需求正在爆发式增长。根据国家数据局披露的数据，中国日均词元调用量，在短短两年时间里，从2024年初的1000亿，飙升到了2026年3月的140万亿。这是什么概念？两年时间，增长了超过1400倍！

这个数字有多惊人？我们可以做个对比。2003年，中国全年的手机短信发送量才刚刚突破2000亿条（14万亿个字）。而现在，我们一天的词元调用量就达到了140万亿次。这意味着，AI应用的普及速度已经远远超过了我们的想象。

根据行业的预测，2026年，中国词元经济相关的产业规模将突破8000亿元。到2030年，这个数字将冲击3万亿元大关，年复合增长率超过50%。这是一个什么样的市场？这是一个比互联网时代的流量市场还要庞大得多的市场。

在这个市场里，词元就像石油一样，成为整个时代的核心原材料。谁掌握了词元的生产，谁就掌握了这个时代的财富密码。

就像在石油时代，洛克菲勒通过控制石油的开采、冶炼和运输，建立了标准石油帝国，成为那个时代的世界首富；就像在互联网时代，扎克伯

格通过掌握流量，建立了Facebook帝国，成为全球最年轻的千亿富豪。在AI时代，谁能掌握词元的生产和分配，谁就能成为下一个时代的巨头。

这不是危言耸听，这是正在发生的现实。我们已经看到，阿里巴巴迅速成立了由CEO亲自挂帅的“Alibaba Token Hub”事业群，目标就是全面掌控词元的创造、输送和应用生态。字节跳动、腾讯、华为这些巨头，也都在疯狂地布局词元生产线，抢占这个新赛道的制高点。

因为大家都清楚，在AI时代，词元就是新的货币、新的“石油”、新的流量。谁先占领了这个高地，谁就能在未来的竞争中占据主动。

第二章 词元经济的三层格局：你在哪一层，决定你赚多少钱

在GTC 2026的主题演讲上，黄仁勋向全场十万名观众描绘了一个全新的图景：“数据中心相当于全天候运转的词元工厂，原材料是数据和电力，产品是词元。”这句话，第一次把词元经济的生产逻辑，清晰地展现在了世人面前。过去，我们总觉得AI是一个很玄乎的东西，是高科技，是黑箱。但黄仁勋告诉我们，不是的。AI的本质，就是一个工厂。这个工厂把电力和数据当成原材料，加工生产词元这个产品。

就像一个纺织厂，把棉花当成原材料，织成布；就像一个炼油厂，把石油当成原材料，炼成汽油。词元工厂就是把电力和数据加工成词元。在这个工厂里，效率是唯一的核心。

黄仁勋还提出了一个全新的指标：每瓦词元数（Tokens per Watt）。这将衡量未来数据中心的收入能力。为什么？因为“在固定功率限制下，谁的每瓦词元数吞吐量最高，谁就拥有最低的生产成本”。

这意味着，词元经济的竞争本质上就是效率的竞争。谁能用一度电生产出更多的词元，谁就能拥有更低的成本，谁就能在市场上拥有定价权。

英伟达公司的技术迭代，其实一直都是围绕着这个目标在走。从CUDA到TensorRT，从A100到H100，再到最新的B200，所有的技术创新本质上都是为了提升词元的生产效率，都是为了让每一度电能产出更多的词元。

黄仁勋举了一个非常生动的例子：Fireworks AI和Lynn两家公司，在没有更换任何硬件的条件下，仅仅靠英伟达更新的软件栈和推理算法，词元的生成速度，就从每秒约700个，提升到了近5000个。硬件一点都没换，只是软件升级了一下，效率就提升了7倍多。这意味着什么？意味着同样的电力，同样的服务器，过去一天能生产100万个词元，现在能生产700万个。成本直接下降了85%以上。

这就是词元生产的魔力。只要有人能提升效率，就能把成本降下来，就能在市场上拥有无与伦比的竞争力。

而在整个词元经济的体系里，按照词元生成和消费的流转过程，我们可

以把所有的人分成三个层次。这三个层次，对应着完全不同的赚钱逻辑，也对应着完全不同的财富高度。

简单来说：你在哪一层，决定了你能赚多少钱。

2.1 下游：词元消费者——用时间换钱，人人可复制

最底层的，是词元的消费者。

什么是词元消费者？就是那些购买词元、消耗词元，用AI工具来干活的人。

有人可能会问，我用ChatGPT写文案，用Midjourney画图，用AI做PPT，我不就是在使用AI吗？没错，人们是在使用AI，但本质上，是词元的消费者。

消费者调用大模型的API，消耗了模型生产出来的词元。消耗这些词元完成工作，然后把工作成果卖给客户，赚一点辛苦钱。

这就是下游的逻辑。

在这个层次里，他的商业模式是什么？是调用大模型→消费词元→出卖服务。比如，现在网上有很多人用AI帮别人写文案、做PPT、剪视频，接单赚钱。他们的成本是什么？是调用AI的费用，也就是词元的费用。他们的收入是什么？是客户给他们的服务费。

听起来好像不错。但问题是，这个模式有一个致命的缺陷：它是可复制的，是没有壁垒的。

首先，你的服务是完全可替代的。客户找你写文案，明天他自己学会了用豆包，一分钟就能写完，为什么还要找你？或者，有另一个人愿意以更低的价格接这个单，你怎么办？因为大家用的都是同一个大模型，消耗的都是同样的通用词元，你们提供的服务质量未必有本质区别。唯一的竞争维度就是价格。

因为AI工具是所有人都能用的。你会用ChatGPT写文案，别人也会用。你会用Midjourney画图，别人也会用。这个门槛太低了，低到只要会打字，就能干。那结果是什么？就是内卷。别人收100元，我收80元，他

收50元，最后价格越来越低，利润越来越薄。人们只能靠拼时间、拼体力来接更多的单，才能赚更多的钱。

其次，你的收入是线性的，而且有天花板。你一天最多工作8个小时，最多接10个单，赚1000元钱。你不可能一天工作24小时，也不可能无限地接单子。因为本质上你出卖的，还是你自己的时间。哪怕你把效率提高了10倍，你一天最多也只能接100个单，赚1万元钱。这就是线性增长的天花板。

最后，你没有任何积累。你今天接了一个单，做完了，钱拿到了，这个单就结束了。它不会给你带来任何长期的资产。你明天如果不接单，就没有收入。你所有的财富，都来自你当下的劳动。一旦你停止劳动，财富的流入就立刻停止。

随着AI工具越来越便宜、越来越好用，这些收入空间会越来越小。因为当AI工具足够便宜、足够好用的时候，客户为什么要找你？客户自己直接用AI不就行了？

比如，过去可以帮客户写公众号文章，一篇收500元钱。现在客户自己用ChatGPT，花几分钱的词元费，几分钟就写出来了，可能比你写得还好。那这个服务，还有什么价值？

这就是为什么我们说，这是低端的赚钱方式。它本质上，还是在消费词元，而不是生产词元，只是把上游生产好的通用词元转手包装了一下，卖给了终端客户。你在整个产业链里，处于最下游的位置，赚的是最微薄的辛苦钱。

这就像什么？就像工业时代的流水线工人。工厂生产出了机器，工人用这个机器，帮工厂加工零件，工人赚的是计件工资。工人干得多，赚得多，但很难发财，因为其没有掌握生产资料，只是在消耗别人的生产资料，出卖自己的时间。

在词元经济里，下游的消费者就是这样。他们没有词元的生产线，只能买别人的词元来用。他们赚的是最底层的、最辛苦的、最少的钱。

AI工具使用者往往在追着技术跑，常做最底层的、可替代的工作，很难赚到大钱。这就是下游的困境。

2.2 中游：词元服务商——赚差价的中间商，惨烈内卷

比下游高一个层次的，是词元的服务商，也就是中游的玩家。

什么是词元服务商？就是向上游采购词元，然后打包成各种服务，卖给下游的用户，赚取中间的差价。

比如，很多AI应用商店整合了很多大模型的API，然后给用户提供一个统一的入口。用户在这里买词元，比直接去OpenAI买要便宜一点，或者更方便一点。词元服务商赚的就是这个差价。

再比如，很多SaaS公司把AI能力集成到自己的软件里，然后给企业提供服务。他们的成本，就是向上游买词元的费用；他们的收入，就是企业给他们的软件服务费。本质上，他们是在赚词元的差价，利用智能生产力扩散的信息差赚钱。

还有那些做AI代理的，帮大模型做推广，拉客户，然后拿返点。这本质上也是中间商，也是赚差价。

这个模式听起来好像比下游好一点。毕竟，不用自己干体力活了，中游做的是服务，是流量生意。

但问题是，中游的门槛同样很低，竞争同样惨烈。

因为词元这个产品是标准化的。OpenAI的词元和Anthropic的词元，本质上没有太大的区别。你能卖，我也能卖；你能拿到返点，我也能拿到。

那结果是什么？又是价格战。有人把词元降价10%，我降价20%，他降价30%。最后，利润被压得越来越薄，很多中间商甚至赚不到钱，只能靠走量。

而且，上游的巨头随时都可能把中游干掉。比如，OpenAI自己就有官方的API，它为什么要让中间商来赚这个差价？当市场成熟了，它完全可以直接对接用户，把中间商给绕过去。就像我们过去看到的，很多电商平台的品牌代理商，最后都被品牌方自己的旗舰店给干掉了。很多手机的经销商最后都被品牌自己的线上商城给“干掉”了。

在词元经济里，这个趋势会更明显。因为上游的巨头拥有算力、拥有模型，自己就能直接触达用户。中间商只是在市场早期存在，帮助他们做推广、做流量。当市场成熟了，这些中间商的价值，就会越来越小。

所以，中游的玩家看起来好像比下游高端一点，但本质上还是在给上游打工。他们没有定价权，没有壁垒，只能在巨头的缝隙里辛苦地赚一点差价。而且，这个市场的竞争比下游还要惨烈，因为所有的玩家都在拼价格、拼流量，最后活下来的没几个。

2.3 上游：词元生产者——掌握定价权，赚长期的大钱

整个词元经济体系里最顶层、最赚钱的，就是上游的词元生产者。

什么是词元生产者？就是那些能自己生产词元、供给词元、定义词元价值的人。

他们拥有自己的词元工厂，拥有自己的算力，拥有自己的模型。他们把电力和数据加工成词元，然后卖给整个市场。他们才是整个产业链的源头，才是真正掌握生产资料的人。

这个层次的商业模式是什么？是建立词元工厂→生产词元→供给市场。不再是买别人的词元来用，自己就是词元的卖家；不再是给别人打工，自己就是规则的制定者。

这意味着什么？意味着你拥有了定价权。

因为词元是整个AI时代的刚需，所有人都需要它。你生产出来的词元，整个市场都会来买。你可以决定它的价格，决定它的标准，决定整个行业的游戏规则。就像石油时代的石油公司，不管经济怎么波动，不管油价怎么涨，它们永远都是最赚钱的。因为所有人都需要石油，普通人没得选。

词元生产者也是一样。不管AI应用怎么变，不管下游的模式怎么变，大家都需要词元。而提供词元的人只要把生产线建好，就能源源不断地赚钱。

而且，这个生意有一个非常恐怖的特点：边际成本趋近于零。

什么意思？就是在搭建词元生产线时，只需要投入一次成本。买服务器、建机房、调模型，这些是一次性的投入。建好之后只要交电费，就能源源不断地生产词元，销往市场。

生产的词元越多，平均成本就越低。因为固定成本已经摊薄了。卖出去的每一个词元，几乎都是纯利润。

这就是为什么，那些上游的巨头，都在疯狂地布局词元生产线。因为他们知道，这才是真正的“印钞机”。

比如MiniMax这样的中国AI公司就是典型的上游词元生产者。他们搭建了自己的词元生产线，把词元的价格降到了极致。其调用价格只有0.3美元每百万词元，是Claude Opus的1/17。

凭借这个极致的性价比，他们的模型在OpenRouter这个全球平台上登顶全球调用量榜首。全球的开发者都来买他们的词元。他们不需要去做下游的应用，不需要去做中游的服务，只要卖词元就能赚得盆满钵满。

还有字节跳动、阿里这些巨头，他们在内蒙古、宁夏等“东数西算”枢纽节点，搭建了大规模的词元工厂。他们利用西部便宜的绿电，利用液冷的技术，把词元的生产成本压到了最低，然后把这些词元卖给全球的客户。

甚至不仅仅是巨头，普通人也有机会进入上游。过去我们觉得，建数据中心、搞大模型，那是巨头才能干的事。但现在技术的发展已经把这个门槛降得很低了。

人们不需要买几十亿的服务器，可以采用租赁的方式拿到算力；不需要自己从零训练大模型，可以用开源的模型，做微调，做优化；不需要自己建机房，可以用云服务，用边缘计算。

只要掌握了词元生产的技术，只要能把效率提上去，把成本降下来，每个人都能搭建自己的小词元工厂，都能自己生产词元，卖给市场。

这就是本书要教给大家的东西。我们不是要教大家怎么当下游的服务承接者，怎么用AI工具接单赚钱。我们也不是要教大家怎么做中游的服务商，怎么赚微薄的差价。

我们要带大家直接进入词元经济体系的上游，去做一个词元生产者。我们要教大家怎么搭建属于自己的词元生产线，怎么拥有一台能持续生财的“印钞机”。因为只有这样，我们才能在AI时代真正掌握财富的密码，真正实现财富的自动增长。

2.4 三层格局的本质：生产资料的归属

这三层格局的本质，其实就是生产资料的归属。我们都学过政治经济学，都知道在任何一个经济时代，财富的分配永远都是向掌握生产资料的人倾斜的。

农耕时代，生产资料是土地。拥有土地的地主，永远比没有土地的佃农要富裕得多。佃农再努力，他也只是在给地主打工，很难比地主富有。

工业时代，生产资料是机器和工厂。拥有工厂的资本家，永远比流水线上的工人要富裕得多。工人再努力，他也只是在给资本家打工，很难比资本家富有。

在互联网时代，生产资料是流量和服务器。拥有平台的巨头永远比平台里的商家和创作者要富裕得多。商家再努力，他也只是在给拥有平台的巨头打工，很难比拥有平台的巨头富有。

而在AI时代，生产资料是什么？就是词元的生产线。

拥有词元生产线的生产者，永远比那些消耗词元的消费者、中间商要富裕得多。因为他们掌握了生产资料，掌握了整个时代的核心供给。

下游的消费者没有生产资料，只能出卖自己的时间，所以赚最少的钱。

中游的服务商也没有生产资料，只能倒买倒卖，赚点差价，所以赚得也不多。

只有上游的生产者拥有生产资料，掌握了词元的供给，所以能拥有定价权，能赚长期的大钱。

这就是为什么黄仁勋要反复强调“词元工厂”“每瓦词元数”，因为他知道，这才是这个时代的核心。

现在，大家可以问一下自己：在词元经济的这三层格局里，自己现在在哪一层？未来想要去哪一层？

如果还在下游，每天想着怎么用AI工具接单，那就要小心了，因为那个市场会越来越内卷，越来越没有空间。

如果在中游，做着中间商的生意，那也要小心了，因为上游的巨头一直在想方设法，随时都会把中游的利润侵占掉。

而如果能进入上游，成为一个词元生产者，那就会发现一个全新的世界的大门在自己面前打开了。我们不再需要拼时间，不再需要拼价格，只要把生产线建好，我们就能实现持续盈利，就能实现财富的自动增长。

这正是词元经济的财富增长逻辑。

第三章 词元不等于代币：合规、安全、国家支持的财富赛道

很多人第一次听到“词元”这个词的时候，第一反应就是：这是不是就是那种虚拟货币？是不是就是比特币？是不是又是那种炒作的骗局？毕竟，过去几年我们见过太多打着“区块链”“代币”名义的骗局了。很多人被割韭菜割怕了，所以一听到什么“元”什么“币”，就本能地警惕。

但需要明确的是：词元和那些虚拟货币、代币是两码事。不仅不是一回事，两者之间甚至有着本质的区别。而且，词元是国家大力支持的，是合规的，是安全的，是真正服务于实体经济的财富赛道。

这一点一定要搞清楚，因为这决定了我们能不能放心地、大胆地去布局这个赛道。

3.1 一禁一倡：国家的明确态度

先来看一下，最近国家出台的两个政策，就能明白国家的态度了。

2026年2月，中国人民银行、国家发展改革委等8部门联合发布了《关于进一步防范和处置虚拟货币等相关风险的通知》，同时证监会也发布了监管指引。这个文件里明确说了，严禁境内虚拟货币（也就是我们熟悉的BTC、ETH、稳定币这些非法定代币）的发行、交易和中介，这些全部都是非法的。同时，严管境外的RWA（现实世界资产）的代币化。

这个政策非常明确，就是打击虚拟货币，打击那些炒作的代币。国家的态度很清楚：这些东西是虚的，是骗人的，是会引发金融风险的，我们坚决不要。

与此同时，国家大力推动数字经济与实力基础设施建设，在2026年3月的全国两会上，《政府工作报告》里首次写入了“算电协同”四个字。这标志着算力与电力系统的深度融合，正式上升到了国家战略的层面。

而“算电协同”是什么？这就是词元经济的核心基础。因为词元的生产就是把电力转化为算力，再把算力转化为词元。算电协同，就是为了让这个过程更高效、更便宜、更规模化。

一个禁止，一个倡导。国家的态度已经非常明确了。

对于虚拟货币、代币，国家坚决打击，坚决禁止。对于词元经济，国家大力支持、大力推动，把它上升到国家战略的高度。这两者国家分得清清楚楚，为什么？因为它们虽然看起来都用到了算力，都用到了电力，但它们的本质完全不一样。

3.2 本质区别：一个是炒作的泡沫，一个是创造价值的生产

比特币这种虚拟货币，它的本质是什么？它是一种虚拟资产。它消耗算力和电力的过程，是为了什么？是为了维持它自身的区块链网络，维持它自身的交易和炒作。

简单来说，它挖币的过程，除了产生那个币本身，没有创造任何其他的社会价值。它既不能写文案，也不能做诊断，也不能优化生产流程。它什么都干不了，它唯一的作用，就是拿来炒作，拿来炒价格。而且，它还浪费了大量的能源。过去，比特币挖矿一年消耗的电力，比一个中等国家的用电量还要多。这些电力，没有用来给工厂供电，没有用来给居民供电，消耗在了维持其运行的哈希计算上，只是为了生产出一个虚拟币，拿来炒作。

这就是为什么国家要打击比特币。因为它脱实向虚，不创造任何价值，只会引发金融风险，只会浪费能源。

然而词元的情况完全不一样。

词元消耗的算力和电力，是为了什么？是为了支撑AI处理信息，是为了服务生产生活，是为了创造实实在在的社会价值。

使用者消耗一个词元，AI就能帮他写一篇文章，做一个设计，诊断一个病例，优化一个生产流程。这些都是实实在在的价值，都是能推动经济发展、推动社会进步的。

比如，浪潮云洲和黑猫集团的合作案例，他们把20位老师傅40年的经验，转化成了3.6万条知识图谱，然后训练了一个工业大模型。这个模型，把炭黑的合格率，从82%提升到了94%，一年给公司节省的成本，就超过了2000万元。

这个过程，消耗了词元吗？当然消耗了。然而，这些词元的消耗带来了什么？带来了2000万元的成本节约，带来了生产效率的提升，带来了实实在在的经济效益。

再如中国联通的服装行业大模型，他们把设计制版的周期缩短了80%。过去，设计师要花一个月才能做完的设计，现在只要几天就能做完。这背后也是词元的消耗。但这些消耗，带来了什么？带来了服装行业效率的提升，带来了整个产业的升级。

在医疗领域，AI辅助诊断系统能帮帮助医生快速地分析CT影像，提前发现癌症的病灶，这背后也是词元的消耗。但这些词元的消耗拯救了无数的生命，这是多大的社会价值！

这就是词元和比特币的本质区别。

比特币消耗的电力，除了炒作，什么都没留下。词元消耗的电力，留下了效率的提升、产业的升级和社会的进步，留下了实实在在的价值。这就是为什么国家要大力支持词元经济，要坚决打击虚拟货币。因为一个是脱实向虚的泡沫，一个是服务实体的生产力。

3.3 词元经济：国家战略级的出口赛道

词元经济不仅仅是国内的产业，它还是国家战略级的出口赛道。

过去，中国的出口靠的是什么？是服装、是玩具、是家电、是工业品。我们是世界工厂，我们把我们生产的商品卖给全世界。现在，词元经济给我们带来了一个全新的出口赛道：数字出口。

我们把国内的工业电力转化为算力，再转化为词元，然后把这些词元卖给全世界。这就像我们过去出口石油，而现在我们出口“数字石油”，并且中国在这方面拥有无与伦比的优势。

首先，我们有便宜的绿电。中国的水电、光伏、风电，装机量都是全球第一。我国西部的电价非常便宜，一度电只要两毛钱，甚至更低。这就意味着，我们生产词元的成本天生就比别人低。

其次，我们有“东数西算”的战略。我们把算力工厂建在西部的绿电基地，把成本压到了极致。这是其他国家根本比不了的。

最后，我们的技术优化能力全球领先。我们的企业通过算法优化，通过推理加速，能把词元的生产效率提得非常高。这就使得我们生产的词元价格只有海外的1/6，甚至1/10。

这是什么概念？这就意味着，我们的词元在全球市场上拥有碾压性的性价比。现在，我们已经看到了中国的词元正在疯狂地出口。

比如MiniMax的词元在OpenRouter上调用量已经是全球第一了。全球的开发者都来买他们的词元，因为太便宜了，太好用了。

还有阿里云、腾讯云这些云厂商，他们的API服务已经卖到了全球。他们的词元，在东南亚、在中东、在欧洲，都非常受欢迎。

根据统计，现在中国模型在全球词元市场的份额已经和北美平分秋色了，而且这个份额还在快速地增长。

这就是我们的新出口。过去我们出口工业品，现在我们出口数字产品，出口词元。这个市场比过去的工业品市场还要大得多。而且这个出口是完全合规、完全合法的，是国家大力支持的。因为它能给我们带来外汇，能给我们贡献GDP，能提升我们在全球数字经济领域的话语权。

这就是为什么，国家要把词元经济上升到国家战略的高度。因为它是我们中国在AI时代弯道超车的一个绝佳机会。

3.4 个人的机会：合法拥有，合规运营

很多人会问，国家这么支持，那我们普通人能不能参与？我们能不能自己做词元生产者？会不会有什么合规性的问题？

我可以明确地告诉大家，完全可以。词元是个人可以合法拥有、合法运营、合法供给的。它不像虚拟货币，受到严格监管，风险较高。词元是完全合规的，是国家允许，甚至是鼓励的。

为什么？因为词元的生产在本质上就是AI的推理服务，就是云计算服务。这些服务，本来就是合法的，是正常的商业活动。

搭建一个词元生产线，本质上就是提供了一个AI的API服务，可以按用量收费。这和过去开一个网站、提供一个云服务，没有任何区别，都是

正常的商业行为，都是完全合法的。而且国家现在正在鼓励中小企业、鼓励个人参与到算力的建设与AI的服务中来。因为这能推动整个产业的发展，能激活市场的活力。

过去我们觉得只有巨头才能做这个事，但现在，技术的发展已经把门槛降得很低了。人们不需要有几十亿的资本，不需要有顶尖的技术团队，只要掌握了方法，就能搭建自己的小词元工厂，就能合法地生产词元，卖给市场。

比如，现在有很多个人开发者基于开源的模型，做了一个垂直领域的小模型，然后把它封装成API，卖给这个行业的企业。他们的词元因为是专业领域的词元，所以价格能卖得很高，利润非常可观。

还有一些人利用西部的便宜电力，搭建了小型的推理集群，然后给一些中小AI公司，提供算力服务，提供词元服务。由于成本很低，所以服务很有竞争力，赚了不少钱。

3.5 消除误解：词元不是炒作，是实实在在的生产力

最后，我要再消除一个普遍的误解：词元经济是不是又是一种概念炒作？是不是又是资本割韭菜的工具？真的不是。

词元经济不是炒作，它是实实在在的生产力，是实实在在的产业。日均140万亿的调用量，这不是假的，这是真实的需求。企业需要AI，个人需要AI，这些需求是真实存在的，不是炒出来的。8000亿的产业规模，3万亿的未来市场，这也不是假的，这是实实在在的产业增长，是整个AI产业商业化的必然结果。

现在已经有无数的企业在靠词元赚钱了；已经有无数的开发者在靠词元生存了。这不是泡沫，这是正在发生的商业现实。

词元经济的发展已经给我们的社会带来了实实在在的好处。它让AI的成本越来越低，让越来越多的人能用得起AI。它让千行百业的效率越来越高，让我们的经济越来越有活力。

这就是为什么国家要大力支持它，为什么巨头要疯狂布局它，为什么越来越多的人要投身到这个赛道里来。它不是泡沫，它是未来，它是AI时代真正的财富机遇。

在这里，人们不用去碰那些非法的东西，不用去玩那些炒作的把戏，只要踏踏实实地搭建词元生产线，踏踏实实地生产词元，供给市场，就能合法地、安全地、持续地赚钱。

这正是词元经济给我们每一个普通人带来的最大的机遇。

第二部分

词元经济铁三角：电力—算力—词元

当我们在谈论AI时代经济的新特征时，很多人第一时间想到的是科技感的AI应用、能思考的大模型，或者是一夜爆红的某个智能体应用。但如果我们剥开这层光鲜的外衣，深入整个产业的最底层，就会发现一个无比朴素却又无比震撼的事实：所有的智能、所有的价值、所有的财富，最终都源于一个最基础的物理转化过程——电力，通过算力的加工，最终转化为词元。

这不是一个抽象的技术流程，而是一条实实在在的工业生产线。黄仁勋的话很直白，“数据中心就是全天候运转的词元工厂，原材料是数据和电力，产品是词元。”在这个工厂里，电力是燃料，算力是机器，而词元，就是这个工厂生产出来的、可以在全球范围内交易的大宗商品。

这三者，构成了词元经济牢不可破的铁三角。

如果说词元是AI的“信息积木”，那么算力就是搭建这些积木的“力量”——没有算力，词元就只是一堆零散的碎片，无法被AI理解、分析和重组；没有词元，算力就没有作用对象，再强大的算力，也无法发挥任何价值。

如果说算力是搭建词元积木的“力量”，那么电力就是这种力量的“燃料”——没有电力，算力就会停滞，AI就无法处理任何一个词元；没有算力，电力的消耗就没有意义，那些清洁的风电、光伏，就只能被白白浪费在西部的旷野上。

这个铁三角的运转正在重塑全球的经济格局。过去，我们出口的是衣服、鞋子、家电，是实实在在的工业品；现在，我们出口的是词元，是由中国的电力、中国的算力加工出来的数字大宗商品。一度电，在新疆的戈壁滩上，可能只值0.15元人民币，但当它流过GPU，转化为词元，卖到全球市场，它的价值就变成了几美元，甚至几十美元。这就是AI时代的财富魔法，而支撑这个魔法的，就是电、算力、词元这三个最基础的支柱。

在这部分，我们将深入拆解这个铁三角的每一个环节，告诉你：为什么掌握了电力，就掌握了词元的定价权；为什么推理算力，才是词元财富的真正主战场；那个决定你能否在词元经济中赚大钱的核心公式是什么。

第四章 电力：词元的第一成本，也是第一壁垒

在很长一段时间里，当人们谈论AI的竞争时，讨论的焦点总是芯片的性能、模型的参数、算法的聪明程度。但很少有人意识到，在所有这些光鲜的技术背后，有一个最朴素、最刚性，也最容易被忽略的因素，正在悄无声息地决定着这场全球竞赛的胜负——那就是电力。

对于一个词元工厂来说，电力绝不仅仅是付个电费那么简单。它是整个生产成本的绝对大头，是你能否在激烈的市场竞争中活下来的生命线，更是竞争对手无法轻易复制的最深护城河。

4.1 被忽略的真相：电力才是成本之王

我们先来算一笔账，看看在词元的生产成本里，电力到底占了多大的比重。

很多人以为，买GPU花了那么多钱，算力折旧肯定是最大的成本。但真实的数据会颠覆你的认知。行业运营数据显示，在一个AI数据中心的运营成本中，电力成本的占比高达60%到70%。也就是说，你每花10块钱运营一个词元工厂，就有6—7元钱是用来买电的。剩下的3—4元钱，才是硬件折旧、运维、带宽等其他成本。

这个数字意味着什么？意味着电价的微小波动，都会被成倍地放大到词元的生产成本里。如果你的电价比别人贵一毛钱，那么你的整体运营成本就会贵6—7毛钱。如果你的电价比别人便宜一半，那么你的总成本，一下子就比别人低了30%以上。

这就是为什么黄仁勋会反复强调，数据中心的功率限制是最大的瓶颈。物理法则决定了，一个1吉瓦的工厂不可能凭空变成2吉瓦。电网的输送能力、变电站的容量、当地的能源供给，这些都是硬约束，不是你有钱就能随便加的。在这个固定的功率盘子里，你能生产多少词元，能赚多少钱，首先取决于你每一度电的成本有多低。

我们来看一组真实的对比数据。

在中国东部的核心城市，比如北京、上海、深圳，工业用电的价格是多

少？普遍在0.6—0.8元/千瓦时，核心商业区的商业用电，甚至能突破0.9元/千瓦时。这意味着，如果你把词元工厂建在东部，光是电费这一项，一年就要花费天文数字的钱。

而在西部，情况则完全不同。在内蒙古、宁夏、甘肃这些“东数西算”的枢纽节点，绿电直连的电价是多少？实测数据显示，普遍在0.28—0.36元/千瓦时，部分区域甚至能低到0.28元/千瓦时以下。新疆的电价更低，2026年的机制电价，风电只要0.21元/千瓦时，光伏更是低到0.15元/千瓦时。

这是什么概念？同样一度电，在新疆只要0.15元，在上海要0.9元，差价高达5倍！

这还只是国内的对比。如果我们把目光投向全球，差距就更惊人了。美国同等规模的数据中心，电价是0.7—1.1元/千瓦时。欧洲的电价更贵，很多国家工业电价超过1元人民币。也就是说，中国西部的电价只有东部的三分之一到一半，只有欧美国家的约五分之一到十分之一。

这个差价，直接决定了词元的生产成本。我们来算一下，生产100万个词元，需要消耗多少电？根据国内电力机构的测算，当前主流大模型在高强度推理任务下，生成100万个词元平均耗电量为15—20度电。

如果是在美国，按照最高电价来算，这20度电的成本就是20×1.1=22元人民币。而在新疆，这20度电的成本是多少？20×0.15=3元人民币。光是电费这一项，就差了19元钱！

这还只是电费。如果算上整体的运营成本，中国西部生产一个词元的成本，只有美国的六分之一，甚至十分之一。这就是为什么中国的大模型API能在全球市场上打出极致的性价比。

很多人以为这是中国企业在亏本赚吆喝，但其实不是。因为我们的成本本来就这么低。我们的电费比别人便宜，我们的算力成本比别人低，所以我们就算卖得便宜，依然有利润，而他们如果卖这么便宜，早就亏得底朝天了。

这就是电力的力量。它不是一个可有可无的后勤保障，它是词元生产的第一成本，是决定你定价权的核心要素。谁掌握了更低的电价，谁就能在全球词元市场上，拥有无可匹敌的成本优势，谁就能掌握定价权。

我们可以看看那个在网络上刷屏的“一度电的逆袭”的故事，就能更直观地感受到这种价值的跃迁。

在新疆的戈壁滩上，一度光伏电的上网价格只有0.15元人民币。这度电，如果只是卖给当地的居民，它就只值0.15元。如果把它通过特高压电网输送到上海，卖给东部的工厂，加上传输损耗，加上过网费，它能卖到0.6元左右，价值翻了3倍多。

但是你有没有想过，如果我们把这一度电留在当地，流过GPU，转化为词元，然后通过光纤网络卖到全球市场，它的价值会变成多少？

根据测算，当前主流的大模型用一度电大概能生产50万个词元。如果是通用词元，我们按每百万词元0.3美元的价格来算，这50万个词元就能卖0.15美元，折合人民币大概1.05元。

这还只是通用词元的价格。如果是专业的垂直词元，价格能到每百万词元3美元，那这50万个词元，就能卖1.5美元，折合人民币大概10.5元！

也就是说，同样的一度电，在新疆本地卖，大约值0.15元；送到上海卖，大约值0.6元；转化为通用词元卖，大约值1.05元；转化为专业词元卖，大约值10元！

价值翻了几十倍，甚至上百倍！

这就是词元经济的魔法。它把西部那些曾经因为用不掉而被浪费的绿电，转化成了可以全球交易的数字商品，实现了价值的指数级跃升。过去，我们出口的是煤炭、石油，是低附加值的能源产品。现在，我们出口的是词元，是经过算力加工的高附加值的数字产品。这就是中国在全球贸易中的全新竞争力。

这还不是终点。随着我们的模型效率越来越高，每度电能生产的词元越来越多，这个价值的差距，还会越来越大。

4.2 中国的王牌：绿电优势，全球独一份

既然电力这么重要，那么中国最大的优势是什么？就是我们拥有全球最丰富、最便宜的绿电资源。

很多人可能不知道，中国的风电、光伏装机量，已经连续多年位居全球第一。我们的西北、西南地区，拥有广袤的土地，丰富的风能、太阳能、水能资源。在过去，这些资源因为本地消纳不了，经常出现“弃风弃光”的现象——发了电，用不完，只能白白浪费掉。

但现在，这些曾经被浪费的绿电突然变成了词元经济时代最宝贵的财富。因为AI数据中心恰恰是最大的用电大户。一个大型的AI数据中心一年的用电量堪比一个中型城市。这些西部的绿电终于找到了最佳的消纳对象。

更重要的是，这些绿电不仅清洁环保，而且极其便宜。

我们来看甘肃庆阳的例子。庆阳是“东数西算”的国家枢纽节点之一。最近，甘肃电投的绿电聚合项目正式并网发电。这个项目整合了风电、光伏、储能，给当地的算力园区供电。建成之后，园区55%的用电都由绿电供应，最终的到户电价不超过0.4元/千瓦时。

这个价格是什么概念？比东部地区的电价便宜了一半。而且这还是绿电，符合国家的“双碳”目标。

在内蒙古情况更是如此。内蒙古的风电、光伏资源极其丰富，当地的绿电直供价格能到0.3元以下。很多头部的科技公司，比如字节跳动、阿里巴巴，都把自己的AI数据中心建在了这里。因为在这里，他们能拿到最便宜的电，能最大限度地降低词元的生产成本。

国家层面更是把这一点上升到了战略高度。2026年3月，国家数据局明确提出，八大“东数西算”枢纽节点的新建算力设施，绿电应用占比必须达到80%以上。这不是一个倡议，这是一条硬性的红线。达不到这个标准，你就拿不到能耗审批，就不能并网运营。

这个政策一下子就把绿电的门槛提起来了。这意味着，未来如果你想建大型的词元工厂，你就必须拿到足够的绿电指标。而那些拥有丰富绿电资源的西部节点，就成了优先的选择。

当然很多人会问，风电、光伏不是不稳定吗？有风的时候有电，没风的时候没电，白天有太阳有电，晚上没电，那我的词元工厂要24小时运转，怎么办？

这其实早就有了解决方案。现在的西部绿电项目，都是“风光储一体化”的。就是风电、光伏，加上储能，加上电网的调节。比如甘肃庆阳的那个项目，就配套了储能系统，在发电多的时候，把电存起来，在发电少的时候，把电放出来，保证算力园区的供电稳定。

而且，AI算力的负载本身就是弹性的。我们可以把那些对实时性要求不高的任务，比如数据预处理、模型微调等任务，放在发电高峰的时候来做。把那些对实时性要求高的推理任务，放在发电低谷的时候来做。通过这种弹性的调度，我们就能最大化地利用绿电，哪怕它是波动的，我们也能很好地利用，一点儿都不浪费。

阿里巴巴在宁夏的数据中心就是这么做的。全部采用当地的风电、光伏供电，配套了大规模的储能系统，还做了弹性的算力调度。最终，绿电占比超过了90%，供电稳定，完全能满足AI数据中心24小时运转的需求。

这就是绿电的优势，它不仅便宜，而且稳定，完全能支撑词元工厂的持续运转。

这就是“东数西算”工程的真正深意。它不仅仅是把数据搬到西部去处理，更是把西部的能源优势，转化为数字经济的竞争优势。过去，我们把西部的电，通过特高压电网输送到东部，给工厂用、居民用，现在，我们把算力工厂搬到西部去，直接在当地把电力转化为算力，然后通过光纤网络，把词元输送到全世界。

这是一种更高效率的资源配置，因为电的传输是有损耗的。特高压的损耗率已经很低了，但还是有4%左右。而词元的传输，几乎是没有损耗的。在西部生产好词元，通过光纤一秒钟就能传到上海、传到纽约、传到伦敦，不消耗额外的能源。

这就是中国的王牌。我们拥有全球最大的绿电产能，拥有全球最低的绿电价格，这是其他国家根本比不了的。欧洲能源危机之后，电价涨得离谱；美国的通胀高企，工业用电价格一路飙升。而在中国，我们有西部源源不断的低价绿电，这就给我们的词元工厂，提供了源源不断的低成本能源。

4.3 电力壁垒：为什么这是别人抄不走的护城河

很多人会问，电力这个东西，不就是买电吗？有钱谁不能买？为什么说它是壁垒？

如果你这么想，那就太天真了。电力的壁垒是天然的，是地理的，是政策的，是别人根本抄不走的。

首先，电力资源是高度地域化的。便宜的绿电只存在于西部的那些特定的枢纽节点。你不可能把内蒙古的风搬到上海去，你也不可能把新疆的太阳能搬到深圳去，这些资源是绑定在土地上的。

而且，数据中心这种高耗能的项目对电网的接入要求极高。一个大型的AI数据中心，动辄几十万千瓦的负荷，不是随便哪个地方的电网都能承载的。西部的那些枢纽节点，国家已经提前规划好了配套的电网，提前建好了特高压的通道，才能支撑这么大的负荷。东部的城市，电网早就满负荷了，你根本接不进这么大的用电负荷，就算你接进去了，电价也贵得离谱。

其次，电力的壁垒是政策的壁垒。国家的“东数西算”规划，已经把算力枢纽的布局定下来了。绿电80%的红线，更是把东部的大部分项目都挡在了门外。未来大型的算力中心，只能建在西部的枢纽节点。这些节点的土地指标、能耗指标、绿电指标都是有限的，你先抢到了，别人就抢不到了。

比如光环新网，这家国内顶尖的IDC服务商，早就在呼和浩特布局了。他们给字节跳动的AIDC提供服务，机柜数量超过8.2万个，早就拿到了当地的低价绿电资源，早就把基础设施建好了。后来者再想进去，不仅拿不到那么便宜的电，连地和电网的接入容量都没了。

这就是护城河。当别人反应过来，想建算力中心的时候，发现最好的资源已经被人抢光了。你只能去那些电价更贵的地方，你的成本天生就比别人高，你根本没法竞争。

我们来看一个真实的案例。字节跳动在乌兰察布建了一个超大规模的AI数据中心，这个数据中心，全部用当地的风电、光伏供电，绿电占比超过90%。当地的电价，只有0.28元/千瓦时。这意味着，字节跳动的算力生产成本天生就比那些建在东部的公司低了30%以上。

这就是为什么字节跳动能以极具竞争力的价格，向外提供算力服务。因

为它的成本底线，比别人的售价还要低。你怎么跟他打价格战？你打价格战，你每卖一个词元亏一分钱，它每卖一个词元赚一分钱。打到最后，你亏光了，它还在赚钱。

这就是电力壁垒的威力。它不是你靠技术就能追上来的，也不是你靠钱就能砸出来的。它是资源的垄断，是地理位置的垄断，是政策红利的垄断。

这个壁垒还在越来越高。随着国家对“双碳”目标的推进，对绿电的要求越来越高。未来，你生产的词元是不是绿电生产的，都会成为一个重要的标准。出口到海外的词元，人家还要查你的碳足迹。如果你用煤电生产算力，可能还要被征收碳关税。而我们用绿电生产的词元，不仅成本低，而且还符合全球的环保标准，在国际市场上更有竞争力。

这就是为什么说，谁掌握了更低的电力成本，谁就掌握了词元的定价权。在词元经济的这场全球竞赛中，电力就是第一道壁垒，也是最高的一道壁垒。

第五章 算力：词元的生产机器

搞定了电力这个燃料，接下来，我们就要来看生产词元的机器——算力。

很多人一提到算力，第一反应就是GPU，就是显卡，就是那些用来“挖矿”、用来玩游戏的芯片。但在词元经济的语境下，算力绝不是一堆显卡的堆砌。它是一个完整的工厂，是一个把电力转化为词元的精密系统。

这个工厂的核心早就已经发生了翻天覆地的变化。过去大家都在抢着训练算力，抢着训练大模型，但现在战场已经彻底转移了。推理算力才是词元财富的真正主战场。

5.1 算力不是显卡，是词元工厂

我们先来纠正一个普遍的误区：算力是显卡。

当然，GPU是算力的核心载体，但如果你把算力理解为买几块GPU插在服务器上，那就太肤浅了。就像你不能把一个工厂就理解为几台机床一样，一个真正的词元工厂是一个完整的系统。

词元并非凭空产生，而是通过消耗大量的计算资源，由大型模型对数据进行处理而生成的。因此，词元生产技术的核心，就是如何以最高效、最低成本的方式，将底层的物理算力（如FLOPs，即浮点运算次数）转化为上层的逻辑词元。这一过程的技术栈极为复杂，涵盖了从硬件到算法的多个层面。

首先，是硬件层。除了GPU芯片，还有服务器、高速互联网络、散热系统、供电系统，这些东西缺一不可。如果你买了最好的GPU，但你的散热不行，芯片过热降频，那性能就发挥不出来。你的网络不行，GPU之间的数据传不动，那再多的GPU也没法协同工作。

其次，是软件层。这是最容易被忽略的部分。很多人以为，我买了GPU，插上电，就能生产词元了。哪有这么简单？你需要有推理引擎、有调度系统、有模型优化的算法。同样的一块GPU，用不同的软件，能

生产出来的词元数量，可能差了好几倍！

最后，是运营层。你怎么调度这些算力？怎么把不同的任务分配给不同的机器？怎么在负载高的时候弹性扩容？怎么在负载低的时候省电？这些运营能力直接决定了你的工厂的生产效率。

所以，算力是一个完整的词元工厂。它是硬件、软件、运营的结合体。它的目标只有一个：把输入的电力和数据，最高效地转化为输出的词元。

黄仁勋在GTC大会上举了一个特别震撼的例子。Fireworks AI和Lynn这两家公司，他们没有更换任何硬件，就是用原来的那些GPU服务器。仅仅靠英伟达更新了软件栈，优化了推理算法，他们的词元生成速度，就从每秒约700个，提升到了每秒近5000个！

7倍！同样的硬件、同样的电力，只是把软件优化了一下，产出就提升到原来的7倍！

这是什么概念？这意味着，你花1000万买的硬件，原来一天能生产700万个词元，现在不用多花一分钱买硬件，不用多花一分钱买电，一天就能生产4900万个词元。你的单位词元成本一下子就降到了原来的约七分之一！

这就是算力工厂的魔力。它不是简单的硬件堆砌，它是通过软硬件的协同，把每一度电、每一份算力，都压榨到极致，转化为最多的词元。

这也是为什么，英伟达能这么强。因为他们不仅做硬件，还做软件。他们的TensorRT-LLM、他们的Dynamo调度系统、他们的一系列优化库，就是帮你把GPU的性能发挥到极致。同样的GPU，用英伟达的软件栈，能比别人生产更多的词元。

这就是算力工厂的真正含义。它是一个精密的系统，目标就是最大化词元的产出、最小化词元的成本。

5.2 推理算力＞训练算力：词元财富主战场

了解了算力是工厂之后，我们再来看看，这个工厂的需求结构正在发生什么样的巨变。

在过去两年，大家一提到AI算力，想到的都是训练算力，就是那种用来训练大模型的、超大的算力集群。比如训练GPT-4，训练Llama，需要几万张GPU，跑几个月。那时候，大家都在抢训练卡、抢A100、抢H100，因为训练大模型需要这个。

但现在，一切都变了。

当大模型训练完成，正式上线服务之后，真正的消耗就变成了推理。什么是推理？就是用户调用API，输入提示词，模型生成回复的这个过程。每一次对话，每一次词元的生成，都是推理。

而这个需求，有多大？

我们来看一组数据。2026年，全球AI芯片的市场规模，预计突破2800亿美元。其中，推理芯片的市场规模达到了1450亿美元，占比约52%。训练芯片呢？只有950亿美元，占比约34%。

这意味着，推理芯片的市场规模，已经首次超过了训练芯片。

而在中国，这个趋势更明显。IDC的数据显示，2026年，中国的GPU服务器市场，推理算力的占比已经达到了70%，训练算力只占30%。巴克莱银行的预测更夸张，他们说，到2026年，AI推理的计算需求将达到训练需求的4.5倍！

这是什么概念？就是说，现在你花4.5元钱在推理芯片上，才花1元钱在训练芯片上。而且，推理的增速远远超过训练（大模型的计算过程分两个阶段，先是训练过程，得到一个定制的大模型，然后是推理过程，就是响应用户的查询，调用大模型生成回答）。推理芯片市场的年复合增长率超过50%，而训练芯片市场的增速只有20%。

为什么会这样？

因为训练是一次性的。你训练一个大模型，可能花几个月，花几亿美元就搞定了。但是推理是永久的。只要这个模型在运行，只要用户在调用，你就需要源源不断的推理算力来生产词元。

更何况随着AI应用的爆发，推理的需求正在指数级增长。

中国的日均词元调用量，从2024年初的1000亿，涨到了2026年3月的140

万亿。两年增长了1400倍！这些调用量，99%都是推理。每一个词元都是推理算力生产出来的。

更可怕的是智能体的爆发。过去，用户和AI聊天，一次对话可能消耗几千个词元。现在，AI智能体要自动完成一个任务，比如帮你写代码、帮你做数据分析、帮你处理工作，它可能要调用模型几百次，消耗几十万个，甚至几百万个词元。

这就意味着，推理的需求还会继续爆炸增长。黄仁勋说，过去两年，推理的计算需求增长了1万倍。这个拐点才刚刚到来。

这就是为什么现在所有的巨头都在疯狂地布局推理算力。头部厂商新增的算力，70%都投向了推理场景。算力租赁市场中，推理算力的增速达到了85%，远远超过了训练算力的35%。

因为大家都明白了，未来的财富不在训练里，而在推理里，在词元的持续生产里。

黄仁勋在GTC大会上还专门提到了一家公司，CoreWeave。这是一家AI原生的云厂商，他们的核心业务，就是做推理算力，做词元工厂。他们的增长速度有多快？他们的客户，包括OpenAI、Anthropic这些顶级的AI公司，他们的业务，每年都在翻倍增长。

为什么？因为这些大模型公司训练完模型之后，需要大量的推理算力来生产词元，来给用户提供服务。他们自己建数据中心太慢了，成本太高了。所以，他们就把推理的任务交给CoreWeave这样的专业厂商。CoreWeave帮他们建好了词元工厂，帮他们把效率做到极致，他们只需要按调用量付钱就行。

这就是推理市场的分工。专业的人做专业的事。做模型的专注做模型就行，推理的算力、词元的生产，交给专业的词元工厂来做。

推理市场还给了国产芯片巨大的机会。过去，训练大模型必须用最好的英伟达的GPU，因为要极致的性能、要极致的精度。但是推理不一样，推理对精度的要求没那么高，很多场景下国产的芯片性能足够用。

比如华为的昇腾910B芯片，在推理场景下性能已经非常接近英伟达的A100了，但是价格比英伟达的A100低了30%左右，而且配套的软件栈

也越来越成熟。现在很多做词元生产的公司都开始大量采购昇腾的芯片来降低成本。

比如某家国内的AI厂商，他们用昇腾的芯片搭建了一个推理集群。他们的每瓦词元数做到了和英伟达的芯片差不多的水平，但是整体的成本比用英伟达的低了25%。这就意味着，他们的词元能卖得更便宜，更有竞争力。

这就是推理算力的魅力，它不再是英伟达一家独大的市场。国产芯片在这里也能找到自己的位置，能帮我们把词元的成本压得更低。

我们来看一个例子。开鸿智谷，他们做了端云协同的推理方案，把一部分推理任务放在端侧来做，剩下的再放到云端。通过这种方式，他们把词元的消耗降低了40%—70%。他们的AIBOX边缘推理设备甚至能做到零云端词元费。

这就是推理优化的价值。通过优化推理的效率，他们能生产更多的词元，用更少的算力满足同样的需求。这就是推理算力市场的机会。

过去，大家都在抢高端的训练GPU、抢H100、抢H200，但现在，推理算力的市场给了国产芯片巨大的机会。华为的昇腾、寒武纪的芯片，在推理场景下性能够用，而且价格比英伟达的芯片低20%—30%，这就使得很多做词元生产的公司开始大量采购国产的推理芯片，以降低成本。

这就是推理算力的魅力。它不再是少数巨头的游戏，它是一个更大、更广阔、更长期的市场，是词元经济真正的主战场。

我们再来说说液冷这项技术对提升每瓦词元数有多大作用。

过去，GPU都是用风冷来散热的。风冷的PUE，也就是能源使用效率，最好的情况能做到1.4。什么意思？就是你花1.4度电，其中只有1度电，是真正用来给GPU计算的，剩下的0.4度电，都用来散热、给风扇供电了。

这就意味着，你的总电力有差不多30%都浪费在散热上了，根本没用来生产词元。

但是液冷不一样。液冷，就是用液体直接给芯片散热。这种方式散热效

率极高。PUE能做到1.1，甚至1.05。

我们来算一笔账。假设你有一个100兆瓦的数据中心，这是电网给你的总功率，你改不了。

如果是风冷，PUE1.4，那么你真正能用来计算的功率是多少？
100/1.4≈71.4兆瓦。剩下的28.6兆瓦，都用来散热了。

如果是液冷，PUE1.1，那么你真正能用来计算的功率是多少？
100/1.1≈90.9兆瓦。

哦！同样的总电力、同样的电网容量、同样的场地，液冷比风冷多出了19.5兆瓦的计算功率！

这19.5兆瓦，是什么概念？如果你的每瓦词元数是5，那么你一天就能多生产19.5×10^6瓦×5个/瓦/秒×86400秒=842.4亿个词元！

按通用词元每百万0.3美元的价格，这一天，你就多赚了842.4亿/100万×0.3美元=25272美元，折合人民币17万！

一个月，就是约500万元人民币！一年，就是约6000万元人民币！

这还只是通用词元的价格，如果是专业词元，那就是数亿美元的收入！

这就是液冷的魔力。它没有增加你的总功率，没有增加你的电网容量，只是把原来浪费在散热上的电省下来，用来计算、用来生产词元，就直接给你带来了每年数千万甚至数亿美元的额外收入。

这就是为什么现在所有的AI数据中心，都在疯狂地上液冷。因为它能直接提升你的每瓦词元数，直接提升你的收入。

过去，我们建数据中心觉得液冷贵，舍不得上。现在，在词元经济时代，液冷的那点额外成本和它带来的收入比起来根本不算什么。你上了液冷，一年多赚数亿美元，这点投入几个月就回来了。

5.3 核心指标：每瓦电，能产出多少词元

了解了推理算力之后，我们来看一个全新的指标，一个黄仁勋反复强调的，决定词元工厂竞争力的核心指标——每瓦词元数

（Tokens per Watt）。

这个指标是什么意思？很简单，就是每消耗一瓦的电力，你能生产出多少个词元。

黄仁勋说，这将是未来衡量数据中心收入能力的核心指标。他甚至给出了一个公式：收入=每瓦词元数×可用千兆瓦数。

为什么这个指标这么重要？

因为我们之前说过，数据中心的功率是有硬上限的。物理法则决定了你那个地方，最多只能给你1吉瓦的电，不可能再多了。电网的容量就这么大，变电站的容量就这么大，你不可能凭空变出更多的电来。

所以，你的收入不是由你有多少GPU决定的，也不是由你有多大的厂房决定的。它是由这两个数的乘积决定的：你最多能用的电量，乘以你每度电能生产的词元数。

可用的千兆瓦数是固定的，是物理限制，你改不了。那你能改的是什么？就是每瓦词元数。

这个数越高，你同样的电力，能生产的词元就越多，你的收入就越高，你的成本就越低。

这就是所有词元工厂竞争的核心。大家都在同一个功率限制下，谁能把每瓦词元数做得更高，谁就能赢。

我们来算一下这个账。假设你有一个100兆瓦的数据中心。如果你的每瓦词元数是10，那你一天能生产多少词元？100兆瓦，就是1亿瓦。一天24小时，86400秒。那就是1亿×10×86400=86.4万亿个词元。

如果你的每瓦词元数能做到50呢？那你一天就能生产432万亿个词元。同样的电、同样的场地、同样的电网容量，产出翻了5倍！

这就是差距。这就是为什么Fireworks AI的那个案例那么重要。他们把词元的生成从每秒700个提升到每秒5000个，本质上就是把每瓦词元数，提升了约7倍！

同样的GPU，同样的电，原来每瓦能生产约1个词元，现在能生产约6

个。这就意味着，你的收入一下子就翻了约数倍。

这就是为什么，现在所有的公司，都在疯狂地优化这个指标。

怎么优化？

首先，是硬件的效率。新的GPU架构，比如英伟达的Blackwell，比如NVFP4，就是为了提升推理的能效。在不损失精度的情况下，大幅提升每瓦的性能。还有液冷技术，过去风冷的PUE，可能做到1.4，意味着你给GPU用1度电，总耗电是1.4度，其中0.4度用于散热。现在液冷的PUE，能做到1.1，甚至1.05。这意味着同样的总电力，你能把更多的电用在计算上，而不是散热上。这就直接提升了每瓦词元数。

然后，是软件的优化。TensorRT-LLM、模型量化、知识蒸馏、动态批处理，这些技术都是为了让同样的硬件能处理更多的任务，生成更多的词元。比如，把模型从16位量化到4位，精度损失很小，但计算量和内存占用都大幅下降，这就意味着，你同样的GPU一次能处理更多的请求，生成更多的词元。

还有调度的优化。通过智能的调度系统，把GPU的利用率拉满。过去，很多GPU大部分时间都是空闲的，因为请求是突发的。现在通过动态调度，把不同的任务打包处理，把GPU的利用率从30%提升到70%，甚至90%。这就意味着同样的硬件、同样的电力，你能处理更多的请求、生产更多的词元。

这就是每瓦词元数的意义。它把所有的技术优化、所有的效率提升，都浓缩到了这一个指标里。

过去，我们衡量数据中心的指标，是PUE，是能耗效率。现在，PUE已经不够了。我们需要的是每瓦词元数。因为PUE只反映数据中心的整体能耗效率，无法体现算力的产出。但每瓦词元数告诉你，你用这些电，创造了多少收入。

这就是词元经济时代的全新KPI。对于一个词元工厂来说，你的每瓦词元数越高，你的竞争力就越强。你就能用同样的成本，生产更多的词元，你就能在市场上，打出更低的价格，抢走更多的客户，赚更多的钱。

这就是算力的终极意义。它不是用来炫技的，不是用来比谁的参数大的。它是用来把电力最高效地转化为词元，转化为财富的。

第六章 词元生产公式：决定你能否赚大钱

好了，我们有了电力，有了算力，现在，我们终于可以来回答那个最核心的问题了：到底怎么生产词元，才能赚大钱？

这里我们有两个核心公式。这两个公式就是词元经济的财富密码。搞懂了这两个公式，你就知道，你该怎么布局，怎么赚钱，怎么在这个新的经济体系里占据最有利的位置。

这两个公式就是：

词元成本=电力成本+算力折旧+模型效率

词元收入=专业价值+稀缺性+规模效应

而你的利润，就是：

利润=（收入-成本）×规模=低成本生产×高价值定价×规模化供给

这就是词元生产的全部秘密。

6.1 拆解词元成本：每一分钱都要花在刀刃上

我们先来看成本端。词元的成本，到底是由什么构成的？

很多人以为，词元的成本就是买GPU的钱。不对。我们之前说过，词元的成本，是三个部分的加总：电力成本、算力折旧，还有模型效率。

我们一个一个来拆解。

首先，是电力成本。这个我们之前已经讲得很清楚了。它是运营成本的大头，占了60%—70%。这部分的成本，高度依赖于你所在的位置。如果你在新疆，用0.15元的光伏电，你的电力成本就极低。如果你在纽约，用1块钱的工业电，你的电力成本就极高。

这部分是你选址的时候就基本决定了的。选对了地方，你的成本天生就比别人低一大截。

然后是算力折旧。这部分就是你买GPU、买服务器的钱，分摊到每一个词元上的成本。

比如，你买一块H100，花了3万美元。这块GPU的寿命，大概是3—5年。在这几年里，它能生产多少词元，这些买GPU的钱，就分摊到这些词元上。

所以，这部分的成本取决于两个因素：一个是你买硬件的价格，另一个是你硬件的利用率。

如果你能批量采购，拿到更低的价格，那你的折旧成本就低。比如字节、阿里这样的巨头，他们一次采购几万张GPU，价格肯定比小公司买一张、两张要便宜得多。

更重要的是利用率。如果你能把GPU的利用率拉满，让它24小时不间断地生产词元，那同样的折旧成本就能分摊到更多的词元上，单位词元的折旧成本就低。如果你的GPU大部分时间都在闲着，那折旧成本就很高。

这就是为什么，规模化的词元工厂，成本能做得这么低。他们能把硬件的利用率拉到90%以上，24小时不停转。而小公司，可能只有30%的利用率，那成本自然就高了。

最后，是模型效率。这是最有技术含量的一部分，也是最能拉开差距的一部分。

同样的硬件、同样的电力，不同的模型，能生产出来的词元数量，可能差了好几倍。

比如，DeepSeek的模型，他们通过架构优化，推理效率做得极高。他们的单位词元成本，只有GPT的5%。什么概念？同样的GPU、同样的电，OpenAI生产1个词元的成本，DeepSeek能生产20个！

这就是模型效率的威力。

怎么提升模型效率？

比如，模型量化。把模型的权重，从16位浮点数，压缩到8位，甚至4位整数。这样，计算量就小了，内存占用也小了，同样的GPU一次能处理

更多的请求，生成更多的词元。而且精度损失很小，用户根本感觉不到。

比如，知识蒸馏。用一个大模型，教一个小模型。小模型的参数只有大模型的十分之一，但性能差不多。那推理的成本，就直接降到了十分之一。

比如，MoE架构，也就是混合专家模型。这种模型，每次处理一个请求，只激活一部分专家网络，而不是全部参数。这样，虽然模型的总参数很大，但每次推理的计算量很小，效率极高。

还有，推理算法的优化。比如我们之前说的，动态批处理、前缀缓存，这些技术，都能大幅提升推理的效率。

我们来看一个真实的案例。MiniMax，他们的模型，就是通过极致的模型效率优化，加上西部的低价电力，加上规模化的算力折旧摊薄，把词元的生产成本降到了极致。所以，他们能把输入价格，降到每百万词元0.3美元。这个价格比OpenAI的GPT-3.5还要便宜，是Claude Opus的1/17。但他们依然有钱赚，因为他们的成本本来就这么低。

这就是成本端的三个要素。电力，决定了你的燃料成本；算力折旧，决定了你的机器成本；模型效率，决定了你的机器的生产效率。这三个加起来，就是你的词元成本。

谁能把这三个做到极致，谁就能拥有最低的成本，谁就能在市场上拥有无可匹敌的竞争力。

6.2 重构词元收入：从卖苦力到卖价值

搞定了成本，我们再来看收入端。很多人以为，词元不就是按个儿卖吗？一个词元一分钱，卖得越多赚得越多。

不对。词元的价格不是统一的。不同的词元，价格能差100倍，甚至1000倍。

为什么？因为词元的收入不是由数量决定的，而是由它的价值决定的。

词元收入=专业价值+稀缺性+规模效应。

我们一个一个来看。

首先，是专业价值。这是最大的溢价来源。

通用的词元，就是那种普通大模型生产的，用来聊天的，用来写文案的词元，价格很低。比如MiniMax的每百万词元是0.3美元，OpenAI的GPT-3.5是1美元左右。

但是，垂直领域的专业词元，价格能翻10倍，甚至100倍。

为什么？因为这些词元包含了专业的知识，能解决专业的问题，能给企业带来实实在在的收益。

我们来看一个案例，浪潮云洲和黑猫集团合作的工业大模型。黑猫集团是做炭黑的，这是一种工业原料。过去，生产炭黑靠老师傅的经验。20位老师傅，40年的经验，把这些经验提炼成了知识图谱，训练了一个工业专用的大模型。

这个模型生产的词元，就是工业垂直词元。它能精准地控制生产参数，把炭黑的合格率从82%提升到了94%。就这一项，一年就给黑猫集团节省了超过2000万元的成本。

你说，这样的词元，企业愿意付多少钱？

通用的词元，一百万个才几毛钱。但这个工业词元，一百个，企业就愿意付几块钱。因为它能带来实实在在的收益。你用了我的词元，你一年就能省2000万，我收你几十万，你是不是觉得很值？

这就是专业价值的溢价。

再比如，医疗领域的词元。通用的大模型，给你看病，你敢信吗？你不敢。但是，一个经过医疗数据训练的、专业的医疗大模型，它生成的词元，能帮医生看片子、能帮你做诊断、能给你提供专业的医疗建议。这样的词元，企业愿意付100倍的溢价。

还有法律领域的词元、金融领域的词元，都是一样的。这些词元，不是普通的信息积木，它们是经过专业知识精炼过的，是能解决具体问题的，是能创造实实在在的商业价值的。

所以，企业愿意为这些专业词元支付10倍，甚至100倍的溢价。

这就是为什么，同样是生产词元，有的人赚辛苦钱，有的人赚大钱。因为你生产的是通用的大路货，而别人生产的是高价值的专业货。

我们再来看创研股份的案例。他们做了一个营销领域的垂直大模型，这个模型，专门用来做营销内容，做视频生产。他们的模型在同等效果下，词元消耗省了60%。他们还推出了“词元按量包”，专门给营销行业的客户用。他们的词元价格就比通用的高很多，因为客户用了他们的词元，营销的效率能提升3倍。

还有中国联通的那个服装行业的案例，也非常有代表性。他们做了一个元景服装大模型，专门给服装行业用的。这个模型把服装行业的设计知识、制版知识、面料知识，都训练进去了。

过去，服装设计师设计一个新款要花十几天的时间，制版还要花好几天。现在用了这个大模型，设计师只要输入需求，模型就能自动生成设计图，自动生成制版参数，整个周期缩短了五分之四。

这个模型生产的词元，就是服装行业的专业词元。这些词元能帮服装企业把新品的研发周期缩短一大半，能帮他们更快地响应市场的需求。你说，这样的词元，服装企业愿意付多少钱？

因为用了它，一个新款就能提前一周上市，就能多卖几十万元的货，这点成本根本不算什么。

然后是稀缺性。

有些词元之所以贵，是因为它们稀缺。比如，用顶级的H200 GPU生产的超低延迟的词元，用于那些对实时性要求极高的场景，比如自动驾驶、实时翻译。这种词元因为算力资源稀缺，所以价格就高。

还有，那些拥有独家数据、独家模型的词元，也是稀缺的。比如，某个行业的独家数据训练出来的模型，别人没有，那它的词元就有稀缺性，就能卖高价。

最后，是规模效应。

这个我们之前提过，你生产的词元越多，你的成本就越低，你的边际收

益就越高。因为你的固定成本，比如硬件的折旧已经摊薄了。你生产得越多，每多生产一个词元的边际成本就趋近于零。

所以，你的规模越大，你的利润就越高。这就是为什么，头部的词元厂商能越做越大，越做越赚钱。因为规模效应让他们的成本越来越低，竞争力越来越强。

我们来看DeepSeek的案例，就能更直观地感受到模型效率的威力。DeepSeek的开源大模型在全球的开发者社区里非常火。为什么？因为它们的模型性能和GPT-4差不多，但是推理的成本只有GPT的5%。

他们是怎么做到的？就是通过极致的模型架构优化。他们的MoE架构，做得非常高效，每次推理只激活不到10%的参数，但是性能却和全参数激活的模型差不多。这就意味着，同样的GPU、同样的电，他们能生产ChatGPT20倍的词元。

所以他们的API价格极低。输入价格只要每百万词元0.1美元，比MiniMax还要便宜。但是他们依然有钱赚。

这就是模型效率的魔力，它能把你的成本，压到别人无法想象的程度。

6.3 利润公式：两条路径，两种财富逻辑

好了，现在我们把成本和收入合起来，就得到了利润公式：

利润=低成本生产×高价值定价×规模化供给

这个公式告诉我们，在词元经济里，其实有两条完全不同的赚钱路径。

第一条路径，是低成本的规模路径。

就是我们刚才说的，把成本做到极致。拿到最便宜的电力，用最高效的算力，用最优化的模型，把通用词元的成本，压到最低。然后以极具竞争力的价格，规模化地卖给全球的客户，靠规模赚钱。

这就是MiniMax走的路。他们把词元的价格，降到了全球最低，然后靠海量的调用，靠规模赚钱。他们的词元，虽然单个儿的利润很薄，但是量大，一天几万亿次的调用量，加起来利润就非常可观。

他们还把这些词元卖到了全球。因为中国的成本低，所以我们的词元在全球市场上有极强的竞争力。这就是词元出海，把中国的电力优势、算力优势，转化为全球的数字贸易优势。

这是一条路径，适合那些有能力做大规模业务，有能力把成本压到极致的玩家。

第二条路径，是高价值的垂直路径。

就是刚才说的，做垂直领域的专业大模型。不去和别人打价格战，不去做通用的大路货。而是深耕一个行业，把行业的知识提炼成专业的大模型，然后卖给行业里的客户，靠高溢价赚钱。这就是浪潮云洲、科大讯飞走的路。他们不做通用的聊天大模型，他们做工业的、医疗的、教育的专业大模型。这些大模型，价格是通用模型的10倍、100倍，虽然量没那么大，但是利润极高。

比如科大讯飞，他们的星火大模型在教育和医疗领域，做了深度的优化。他们的教育大模型，能帮学生批改作业，能做个性化的学习推荐。他们的医疗大模型，能帮医生做影像诊断。这些大模型，学校和医院愿意付很高的价格来买，因为它们能实实在在地提升效率，降低成本。

这两条路径，没有好坏之分，都是能赚大钱的路。

低成本的路径，靠的是规模，靠的是成本优势，横扫全球的通用市场。

高价值的路径，靠的是专业，靠的是行业壁垒，深耕垂直的高利润市场。

而且，这两条路最终都会走向同一个方向，就是规模效应。当你做到一定规模之后，你的成本会越来越低，你的利润会越来越高。

我们来做一个具体的测算，你就知道，一个大模型工厂到底能赚多少钱。

假设，我们建一个100兆瓦的大模型工厂，建在内蒙古，电价0.3元/千瓦时。

首先，成本端：

一年的总电费：100兆瓦×24小时×365天×0.3元/千瓦时=100000千瓦×8760小时×0.3元/千瓦时=2.628亿元。

算力折旧：我们假设，我们买了10亿的硬件，分5年折旧，一年折旧2亿。

运维成本：一年0.5亿元。

总成本：2.628+2+0.5=5.128亿元。

然后，收入端：

如果我们做通用大模型，价格按每百万词元0.3美元，汇率7,8.76亿度电能生产438万亿词元，收入就是438万亿/100万×0.3×7=9.2亿元。

利润就是9.2-5.128=4.072亿元！

如果我们做专业大模型，价格是每百万词元3美元，那收入就是438万亿/100万×3×7=92亿元。

利润就是92-5.128=86.872亿元！

这就是词元工厂的赚钱能力！

当然，这是理想情况，实际情况会有负载的波动，会有利用率的问题，但是哪怕打个对折，利润也是几十亿、几百亿的级别。

这就是为什么所有人都在疯狂地建词元工厂。

这就是为什么我们说，搭建词元生产线就等于拥有了一台“印钞机”。因为一旦你搭好了，它就能自己运转，自己生产词元，自己赚钱，不需要你再投入太多的精力。

在新疆，一度电是0.15元。当它转化为通用词元，卖到全球，能卖几元钱，价值翻了几十倍。如果它转化为专业的工业词元，能卖几十元钱，价值翻了几百倍。

这就是词元经济的魔力。它把最基础的能源，通过算力的加工，转化为了高价值的数字商品，实现了价值的指数级跃升。

而这一切，都离不开电、算力、词元这个铁三角。没有电，算力就是一堆废铁；没有算力，词元就是空中楼阁；没有词元，电和算力就失去了价值。

这三者缺一不可，共同构成了词元经济的坚实底座，也共同构成了AI时代财富增长的最底层密码。

第三部分

搭建你的第一条词元生产线

前面的内容里，我们已经讲解了词元经济的底层逻辑，分析了高低端词元市场的核心差异，剖析了成本与价值两条截然不同的财富增长路径，也阐释了规模效应与复利效应在词元生产中的终极真相。至此，多数读者已然对搭建专属词元生产线产生了浓厚的兴趣，渴望拥有这台属于AI时代的“印钞机”，在新的产业浪潮中分得属于自己的红利。

但与此同时，多数初次接触词元生产的从业者都会产生相似的疑问：我是否具备搭建生产线的能力？我缺乏专业的技术背景，没有充足的启动资金，也没有巨头级的产业资源，我真的能够搭建属于自己的词元生产线吗？

答案是肯定的。

在词元经济发展的早期阶段，搭建一套完整的大模型生产流水线，往往需要数亿元的前期投入，以及数千人的专业技术团队，这样的门槛，只有字节跳动、阿里巴巴这样的科技巨头才有能力跨越。但时至今日，产业环境已经发生了翻天覆地的变化：开源模型生态的日益成熟、算力租赁模式的全面普及、一键式部署工具的广泛应用，将搭建词元生产线的门槛降低到了前所未有的程度。

ServiceDirect在《2025中小企业AI报告》中指出，全球77%的中小企业已定期使用AI工具，其中32%的企业每周使用至少7种AI产品。他们并不具备巨头企业的资源优势，却凭借灵活的商业模式、极致的成本控制能力，以及垂直领域的深度深耕，已经在AI时代经济中，成功占据了属于自己的生态位，获得了可观的商业回报。

2025年10月，知乎联合魔搭ModelScope发布的行业首份《AI时代开发者生态白皮书》指出：中国AI市场正处于爆发式增长的万亿级机遇期；940万开发者不仅是代码编写者，更成为把AI潜力转化为真实生产力的“超级个体”。

大势如此，所以哪怕你只是一名独立的个人开发者，哪怕你只有数万元的启动资金，哪怕你只有几个人的微型团队，时至今日你也完全有能力搭建属于自己的词元生产线，通过售卖词元获得收益，享受词元经济带来的增长红利。

在这一部分中，我们将以可落地的实践路径为核心，手把手地引导你从

零开始搭建你的第一条词元生产线。我们将从入门阶段的最小可行产品验证，到算力资源的选型与优化、模型的定制与调优，再到服务的部署与商业化运营，一步步拆解每一个环节的操作细节，估算每个阶段的投入成本，以及可预期的产出。当你完成这一部分的阅读，你将具备动手搭建专属词元生产线的全部能力，正式开启你的词元掘金之旅。

第七章 入门：用最小可行产品验证你的模式

很多初次接触词元生产的从业者，往往会陷入一个认知误区：一提到搭建词元生产线，就默认需要先训练一个千亿参数的通用大模型，需要自建大型的数据中心，需要招募数百人的专业技术团队。当他们初步核算投入成本，发现需要数亿元的资金时，便直接选择了放弃，认为这是自己无法触及的领域。

但实际上，这样的认知完全偏离了词元生产的实践逻辑。这就如同想要进入餐饮行业的创业者，不需要在起步阶段就直接打造覆盖全国的连锁餐饮集团，更合理的路径是先开设一家小型的外卖门店，验证自己的菜品是否符合市场需求，是否存在稳定的付费用户。如果验证通过，再逐步扩张规模，如果验证失败，也不会产生过高的沉没成本。

搭建词元生产线的逻辑完全一致。你不需要在起步阶段就投入全部资源打造大规模的生产体系，更合理的路径是先搭建一个最小可行产品（Minimum Viable Product, MVP），用最低的成本验证你的商业想法，验证你的模式可行性，验证市场上是否存在愿意为你的词元付费的用户。当验证通过之后，再逐步放大生产规模，投入更多的资源，这样就能将创业的风险降到最低。

7.1 词元经济下**MVP**的核心逻辑与独特价值

最小可行产品的核心定义，是用最少的资源投入，打造一个具备核心功能的产品，这个产品能够满足目标用户的核心需求，同时能够帮助创业者验证核心的商业逻辑。

与传统互联网行业的MVP相比，词元经济下的MVP具备独特的优势，其试错成本要低得多。在传统互联网领域，创业者要验证一个商业模式，往往需要开发完整的产品功能，搭建用户运营体系，投入大量的推广资源，整个过程往往需要数月的时间，以及数十万元的投入，一旦失败，沉没成本极高。

但在词元经济中，MVP的搭建过程要简单得多。举例来说，如果你想要切入法律领域，打造专业的法律词元产品，你不需要在起步阶段就训练一个千亿参数的法律大模型，也不需要招募上百人的法律与技术团队。

你只需要选择一个开源的7B参数基础模型，用你手中积累的法律领域数据完成简单的微调，之后将其部署为可调用的API，提供给你身边的律所客户进行测试，询问他们是否愿意使用你的产品，是否愿意为此付费。

如果客户给出了肯定的答复，那就说明你的商业模式是成立的。接下来你就可以投入更多的资源，扩大模型的规模，优化服务的质量，拓展更多的用户。如果客户的反馈并不理想，你也仅仅投入了几千元的成本，以及数天的时间，就完成了这个想法的验证，不会产生过高的损失。哪怕失败，你完全可以调整方向，尝试新的赛道。

这正是MVP在词元经济中的核心价值：用极致低廉的成本完成试错与验证，将创业的风险降到最低。对于初次进入词元生产领域的从业者来说，这是整个流程中最为重要的一步。不要在起步阶段就选择all in，而是要坚持小步快跑的原则，先完成模式验证，再进行规模化的投入。

7.2 MVP的成本结构：极致低廉的试错门槛

很多从业者都会关心，搭建这样一个MVP，到底需要多少投入？答案是，数千元的资金就足够完成整个流程，甚至在部分场景下，数百元的投入就可以完成验证。

我们可以对整个流程的成本进行详细的拆解。

1．基础模型成本

你可以选择开源的基础模型，比如Qwen2-7B、Llama3-8B等，这些模型都是完全免费的，你不需要支付任何授权费用。在Hugging Face等开源模型平台上，有数十万的开源模型可供选择，你可以根据自己的赛道需求免费下载使用。

2．算力成本

你不需要购买专属的服务器硬件，只需要租赁云厂商的算力资源即可。如果选择竞价实例，一张A10显卡的租赁成本每天仅为数十元。微调一个7B参数的模型，仅需要2天左右的时间，算力成本仅为一百余元。而部署阶段的成本更低，每月500元左右的算力资源，就足以支撑数千个用户的并发调用需求。

3．工具成本

当前已经有大量成熟的一键微调、一键部署工具，这些工具大多是开源免费的，你不需要自己编写复杂的代码，只需要通过简单的界面操作，就可以完成整个流程。比如当前主流的LLaMA Factory框架，就提供了可视化的操作界面，你只需要上传你的训练数据，点击开始训练，就可以自动完成LoRA微调的全部流程，不需要掌握专业的技术知识。

综合来看，整个MVP的搭建过程，最多只需要数千元的投入，在部分极简的场景下，甚至只需要数百元就可以完成。

这是什么概念？这个成本甚至低于多数AI培训课程的学费。你只需要投入几千元的成本，就可以验证一个规模达到数十亿的市场机会，这样的投入产出比，是任何创业领域都无法比拟的。哪怕最终验证失败，你也不会有过大的损失，仅仅是获得了一次宝贵的行业经验。

7.3 赛道选择：找到属于你的蓝海市场

在MVP的搭建流程中，第一步也是最为核心的一步，就是选择你的赛道。很多从业者在这个环节会陷入误区，想要做通用型的词元产品，试图覆盖所有的用户群体，但实际上，对于中小创业者来说，通用赛道已经被巨头企业垄断，竞争极为激烈，中小玩家很难在这个领域获得生存空间。

更合理的选择是聚焦于垂直细分赛道，选择一个足够小、足够深的领域，做透这个领域的需求，建立自己的竞争壁垒。当前词元经济的市场中，存在大量尚未被满足的垂直需求，这些领域就是中小创业者的蓝海市场。

比如农业领域的植保大模型，针对农业生产场景，能够识别病虫害、提供种植指导的专业大模型，当前通用模型在这个领域的能力还存在明显的不足；比如中医领域的专业词元，能够理解中医理论、提供养生指导的垂直模型，通用模型很难掌握这些专业的知识；比如非遗领域的文化大模型，能够传承非遗技艺、提供相关咨询的专业模型，这些都是尚未被充分开发的赛道。

在选择赛道的时候，你可以从三个维度进行判断。

1．需求的真实性

这个赛道是否存在真实的、未被满足的需求？目标用户是否愿意为这个需求付费？你可以通过行业调研、用户访谈等方式，提前验证这一点，避免进入伪需求的赛道。

2．竞争的充分性

这个赛道是否已经被巨头企业覆盖？是否存在大量的竞争对手？如果一个赛道已经有了成熟的玩家，并且巨头已经入局，那么中小玩家的机会就会很少，你应该选择那些尚未被充分关注的细分领域。

3．资源的匹配度

你是否具备这个赛道的相关资源？比如你是否有这个领域的专业数据，是否有这个领域的用户资源，是否懂这个领域的专业知识？如果你具备这些资源，你就能够更快地搭建产品，建立自己的壁垒。

7.4 数据版权：容易被忽略的合规风险

在搭建MVP的过程中，很多从业者都会忽略一个重要的问题：训练数据的版权风险。很多人会直接收集网络上的公开数据，或者开源项目的数据，用来微调模型，但实际上，这些数据的版权问题，可能会给你带来潜在的法律风险。

比如，在搭建代码词元MVP的时候，需要收集GitHub上的中文开源项目数据。在这个过程中，我们就需要对数据的协议进行筛选。开源代码的协议有很多种，不同的协议对商用的限制是不同的。

1．**MIT**、**Apache 2.0**等宽松协议

这些协议允许你自由地使用、修改、商用这些代码，不需要开放你的衍生作品的源代码，这些数据是安全的，你可以放心使用。

2．**GPL**等强**Copyleft**协议

这些协议要求，如果你修改了这些代码，或者用这些代码开发了衍生作品，你需要开放你整个作品的源代码，如果你做闭源的商用产品，使用

这些数据就会带来版权风险。

所以，在收集训练数据的时候，你需要对数据的来源和协议进行筛选，优先选择那些宽松协议的数据，避免使用那些有强限制协议的数据，这样才能规避潜在的版权风险。同时，你还需要对数据进行清洗，过滤掉那些可能存在侵权内容的数据，比如受版权保护的文档、未经授权的内容等，确保你的训练数据是合规的。

7.5 典型示例：个人开发者的**MVP**成长之路

王强是一名普通的后端开发者，之前在一家互联网公司工作。2024年，他遭遇了行业裁员，进入了待业状态。在家待业的过程中，他发现了一个普遍的痛点，只是之前自己忙于赶项目进度而忽略了：很多国内的程序员，都需要使用AI工具来辅助编码，但是大厂提供的通用API价格过高，而且对中文的支持不够好，对国内常用的代码框架，比如Vue3、Taro等，支持的效果也不够理想。

于是他产生了一个想法：我能不能搭建一条针对中文程序员的代码词元生产线，提供更便宜、更适配国内场景的代码词元？

他没有充足的启动资金，也没有团队，只有他自己一个人，于是他选择用MVP的方式来验证这个想法。

首先，他选择了开源的CodeLlama-7B作为基础模型，这个模型是专门针对代码场景优化的，完全免费，他从Hugging Face上免费下载了模型文件。

接下来，他开始收集训练数据，他从Github上筛选了所有使用MIT、Apache协议的中文开源项目，整理了这些项目的代码和注释，清洗掉了低质量的代码和侵权的内容，最终整理出10万条高质量的中文代码训练数据。

然后，他使用LLaMA Factory这个一键微调工具，用LoRA技术对模型进行了微调。他租赁了阿里云的竞价A10实例，整个训练过程只用了2天时间，花费了300多元的算力成本。

训练完成之后，他用vLLM工具把模型部署成了API，整个部署过程只用了一天时间，他甚至没有自己编写复杂的代码，只是运行了一行官方

的部署命令，就完成了API的搭建。

之后，他把这个API接入了OpenRouter模型市场，定价为每百万词元1美元，这个价格比大厂的5美元低了80%，具有极强的性价比。

做完这些之后，他就去休息了。第二天早上，当他打开后台查看数据的时候，惊讶地发现，已经有数百个用户调用了他的API，他已经获得了几十元的收入。

这个结果验证了他的想法，市场确实存在对这个产品的需求。于是他开始逐步优化产品、优化模型的效果、优化部署的效率，进一步降低成本，把价格降到了每百万词元0.8美元，吸引了更多的用户。

仅仅用了半年时间，他的日均调用量就突破了1亿，月收入超过了10万元。而他依然是一个人在运营这个项目，没有团队，没有办公室。他现在每天只需要花一个小时的时间，维护一下服务器的状态，剩下的时间，他都在全球旅游，陪伴家人，真正实现了被动收入的财富自由。

王强的案例充分说明，搭建词元生产线并没有想象中那么困难。哪怕你只是一个独立的个人开发者，也完全可以做到。你不需要大量的资金，不需要庞大的团队，你只需要一个可行的想法，用MVP完成模式验证，之后逐步扩大规模，就能够获得成功。

7.6 MVP的标准化搭建步骤：手把手落地指南

那么具体来说，MVP要如何搭建？我们整理了一套标准化的落地步骤，你只需要按照这个流程操作，几天时间就可以完成你的MVP搭建。

第一步，选定你的赛道。确定你要做通用词元还是垂直词元，确定你要切入的具体领域，比如代码、法律、医疗、营销等。不要贪大求全，先选择一个小的细分赛道，把这个赛道做深做透。

第二步，选择基础模型。根据你的赛道，选择合适的开源小模型。比如代码场景选择CodeLlama、通用场景选择Qwen2-7B、医疗场景选择Meditron、法律场景选择LawLlama。这些模型都是开源免费的，你可以直接下载使用，同时优先选择Apache 2.0等宽松协议的模型，规避版权风险。

第三步，数据整理与微调。整理你的行业训练数据，清洗掉低质量、侵权的内容，之后用LLaMA Factory等工具，使用LoRA技术完成模型微调。LoRA技术只需要微调模型的少量参数，成本极低，只需要几千条高质量的数据，就可以完成有效的微调，让模型适配你的垂直场景。

第四步，部署API服务。使用vLLM或者Text Generation Inference等一键部署工具，把你的模型部署成可调用的API。你不需要掌握复杂的网络技术，只需要运行一行部署命令，就可以生成和OpenAI兼容的API接口，用户不需要修改任何代码，就可以直接接入使用。

第五步，接入模型市场开始变现。你不需要自己拓展用户，只需要把API接入OpenRouter、Hugging Face或者国内的硅基流动等模型市场，这些平台会为你带来海量的开发者用户。你只需要设置好你的定价，就可以等待用户调用，获得收益。

通过这五个步骤，只需要几天的时间、几千元的投入，就可以完成MVP搭建，验证你的商业模式是否可行。

如果验证通过，你就可以投入更多的资源，扩大模型的规模，增加算力的储备，拓展更多的用户。如果验证失败，你也不会有过大的损失，可以调整方向，尝试新的赛道，重新进行验证。

这就是词元生产的入门路径。不要在起步阶段就追求大规模的投入，先从小规模的验证开始，确认模式可行之后再逐步扩张，这是最低风险、最高效率的创业路径。

第八章 算力：如何用最低的成本拿到最高的算力

当你的MVP验证通过，你需要开始扩大你的生产规模，这个时候，你首先要解决的核心问题，就是算力资源。

算力是词元生产的核心生产资料，没有算力，你就无法完成词元的生产。但是，算力的价格差异极大，同样的算力资源，有的从业者只需要花费1元的成本，有的从业者却需要花费10元。如何用最低的成本，获得最高的算力，是决定你能否在词元生产中获得利润的关键。

黄仁勋曾经明确指出：算力的能效比，是词元生产的核心竞争力。谁能够在同样的电力成本下，生产出更多的词元，谁就能够在这个市场中获胜。

8.1 算力的多元来源：不只是云服务器

很多从业者对算力的认知存在局限，认为算力就是阿里云、腾讯云等云厂商的标准服务器实例。但实际上算力的来源有很多，你可以选择成本最低的来源，将生产成本压缩到极致。

1．西部低价绿电算力

我们之前曾经提到，我国西部的电价要远低于东部地区，因为西部有丰富的风电、光伏等绿电资源，电价成本极低。因此，西部的算力价格，也要比东部低50%以上。

比如内蒙古乌兰察布、呼和浩特以及贵州的部分地区，这些地方的工业电价约为0.3元/千瓦时，而东部地区的电价要达到0.8元/千瓦时以上，差距极为明显。很多词元生产厂商都把自己的生产线搭建在这些地区，就是为了利用当地的低价电力，降低算力成本。

除此之外，绿电算力还有额外的收益：当前很多企业都有碳中和的需求，使用绿电产生的碳积分，还可以进行交易，带来额外的收入。这进一步降低了算力的综合成本。

2．云厂商竞价实例

云厂商会有大量的闲置算力资源，为了提高资源的利用率，他们会把这些闲置资源拿出来，以极低的价格做竞价实例，价格比正常的实例要低50%—80%。比如，正常的A10服务器，一个小时的价格是3元，而竞价实例的价格只有0.5元。

很多人会担心竞价实例的稳定性，因为竞价实例可能会被云厂商收回，如果其他用户出更高的价格，你的实例将会被终止。但是，对于词元推理的场景来说，这个问题完全可以解决。

因为词元推理是无状态的服务，你只需要做一些简单的容错处理，就可以放心地使用竞价实例。用户的请求如果丢失了，只需要重发一次，不会影响用户的使用。

比如，你可以使用K8s的污点容忍机制，把竞价实例节点设置为优先调度节点，当节点被抢占的时候，系统会自动把请求迁移到其他的节点上。同时，你可以在客户端加入重试机制，当请求失败的时候，自动重试，这样就完全不会影响用户的使用体验。

Fireworks AI就是用这种方式，大量使用竞价实例，将自己的算力成本降低了70%以上，这也是他们能够做到极低词元价格的核心原因之一。

3．全球闲置算力整合

全球有大量的闲置算力资源，很多企业的服务器，白天工作的时候使用，晚上就处于闲置状态；很多个人用户的显卡，平时大部分时间也都是闲置的。这些闲置的算力，都可以被整合起来，用来做词元推理。

CoreWeave就是这个领域的开创者，他们整合了全球的闲置GPU资源，搭建了一个算力池，然后把这些算力卖给AI公司，价格比传统的云厂商要低30%到50%。

国内也有很多类似的平台，比如算力银行，你可以把这些闲置的算力租过来，用来做词元生产，成本比云厂商的标准实例要低一半以上。

4．边缘直供电算力

还有一种成本极低的算力，就是边缘算力。开鸿智谷就把推理节点直接搭建在边缘算力设备，搭建在用户的身边，甚至直接搭建在发电厂的旁

边，使用发电厂的直供电，电价可以低到0.2元/千瓦时，比东部的电价低了75%。

他们的AI BOX设备，直接在边缘侧完成推理，不需要消耗云端算力资源，成本比云端算力低90%。对于企业用户来说，这种边缘部署的方式还能够解决数据隐私的问题，数据不需要离开用户的机房，安全性极高。

这些都是可以使用的算力来源，你不需要只依赖云厂商的标准实例，就可以把这些低成本的算力整合起来，将生产成本降到最低。

8.2 算力的选型：推理不是越大越好，性价比才是核心

选好了算力的来源之后，接下来就是硬件的选型。很多从业者会陷入一个误区，认为算力就要用最好的A100、H100这样的高端GPU。但实际上，对于词元推理场景来说，并非硬件越强越好，性价比才是最重要的。我们可以对不同显卡的性价比进行详细的测算，就能够清楚地看到差异。

参照Piotr Mazurek和Felix Gabriel在《LLM推理经济学》中介绍的推演逻辑，假设在运行7B参数模型的场景下，不同显卡的每百万词元成本差异极大。

1．**T4**卡

这是最经典的推理卡，虽然发布时间较早，但是能效比极高。在INT8量化优化之后，T4卡每小时可以处理约250万词元，租赁成本为每月700元，按当前汇率折算，每百万词元的成本约为0.04美元。对于7B、13B模型的推理场景，T4卡的性价比是最高的。

2．**A10**卡

比T4更新一代的推理卡，性能更强，能效比也很高。它可以支持13B、34B模型的推理，每百万词元的成本约为0.035美元，比T4略低，适合中等规模的词元生产。

3．**L4**卡

英伟达的新一代推理卡，能效比是A10的2倍。黄仁勋在演讲中提到，L4的每瓦词元数是A100的2.5倍，也就是说，同样的电力成本，L4能够生产比A100高2.5倍的词元。折算下来，L4卡的每百万词元成本仅为0.02美元，是当前性价比最高的推理卡。

4．A100/H100

这些是高端训练卡，性能很强，但是价格极高，电费也很高。用来做推理的话，每百万词元的成本要达到0.1美元以上，性价比极低。除非你要运行70B以上的超大模型，否则完全没有必要使用这些高端卡。

根据2026年的行业调研数据，超过80%的词元推理任务都是用L4、A10、T4这些推理卡完成的，使用A100、H100做推理的场景不到5%。因为推理和训练不同，不需要那么强的峰值性能，性价比才是核心的考量因素。

所以对于绝大多数的词元生产线来说，使用L4、A10、T4这些推理卡，比使用高端的A100、H100，性价比要高得多。不要盲目地追求高端硬件，适合你的场景的才是最好的。

8.3 算力的智能调度：把任务分发到最便宜的节点

当你整合了多种不同的算力资源，有的在西部，有的在东部，有的是竞价实例，有的是预留实例，你要如何把这些资源用好？这就需要算力的智能调度。

你可以搭建一个智能调度系统，实时监控所有算力节点的价格、负载、可用性，然后把用户的请求自动分发到当前最便宜的可用节点上。

比如，用户的请求进来之后，系统首先会尝试分发到西部的低价算力节点，如果西部的节点已经满载了，就会分发到竞价实例节点，如果竞价实例也满了，才会分发到正常的标准实例节点。这样，你就能够用最低的成本，处理最多的请求。

Fireworks AI就是这么做的，他们的智能调度系统，能够把用户的请求自动分发到全球最便宜的算力节点上，最终他们的综合算力成本，比其他厂商要低85%，这也是他们能够把词元价格降到极致的核心原因。

除此之外，你还可以实现负载均衡和动态扩缩容，把高峰期的请求分散到不同的节点上，把低峰期的闲置算力利用起来。比如，白天用户的请求多，你就启用所有的算力资源；晚上用户的请求少，你就把一部分节点关掉，节省电费。

通过这样的调度优化，你能够把算力的成本降到极致。很多中小词元厂商，就是靠这样的优化，把成本做到了大厂的几分之一，然后靠性价比优势，抢占了大量的市场份额。

8.4 真实案例：**CoreWeave**的算力优化魔法

CoreWeave，我们之前提到的这个全球首个AI原生云公司，就是靠算力的优化，做到了200亿美元的估值。

刚开始的时候，CoreWeave只是从事GPU租赁业务。他们发现，传统的云厂商算力的价格太高，而且资源的浪费非常严重，很多公司的GPU，大部分时间都是处于闲置状态。于是他们整合了全球的闲置GPU资源，搭建了一个统一的算力池，然后做了智能调度系统，把这些算力整合起来，卖给AI公司。

CoreWeave的算力价格，比AWS要低30%—50%，而且性能还要更好，所以，大量的AI公司，包括OpenAI、Anthropic这些巨头，都选择使用他们的算力。

之后，他们发现很多AI公司需要的不是算力本身，而是词元服务。于是，他们就用自己的低价算力，搭建了自己的词元生产线，生产词元然后卖给用户。他们的词元价格比大厂要低得多，于是迅速占领了市场。

CoreWeave在2025年的营业收入已经超过了50亿美元，他们的综合成本比传统的云厂商要低一半以上。这就是算力优化的力量。这个案例告诉我们，算力是词元生产的核心，谁能够把算力的成本降下来，谁就能够在这个市场中获胜。

第九章 模型：如何选择和优化你的词元生产模型

搞定了算力资源之后，接下来要解决的就是模型的问题。模型是词元生产线的核心，它决定了你的词元的质量，也决定了生产成本。

黄仁勋在演讲中曾提到：“硬件只是基础，真正的魔力来自软件。软件优化，能够把硬件的潜力压榨到极致。我们的Blackwell架构，加上软件的优化，能够把推理的性能提升10倍。”他指出，过去行业的竞争是比谁的模型更大、谁的参数更多。而现在，行业的竞争是比谁的模型更高效、谁的词元成本更低。

综合各方研究报告研判，在过去的几年时间里，开源模型的能力增长极为迅速。2023年，开源模型的能力仅相当于GPT-3的水平；2024年，就达到了GPT-3.5的水平；2025年，已经达到了GPT-4的水平。也就是说，现在的开源模型已经完全能够满足绝大多数场景的需求，你不需要自己训练大模型，使用开源的模型就足够了。

与此同时，过去两年模型的推理成本下降了90%。2024年，一个词元的平均成本是1美元，2026年，已经降到了0.1美元，下降了90%。这背后就是模型优化的力量。

9.1 模型的选择：开源模型已经足够强大

选模型首先要考虑的，是选择开源模型还是闭源模型。

答案非常明确：一定要选择开源模型。闭源模型比如GPT系列，由于无法修改，也无法部署到自己的服务器上，就只能调用别人的API，这样你就只是一个词元的消费者，而不是生产者。所以，如果你想要做词元生产者，就必须选择开源的模型。

现在的开源模型，能力已经非常强大。根据2026年的权威评测，Qwen2、Llama3、DeepSeek这些开源模型，能力已经接近GPT-4的水平，但是它们是完全开源的，你可以免费下载、免费使用、免费部署，这是你做词元生产者的基础。

接下来，你要选择是用通用模型还是垂直模型。

如果你要做通用词元，走成本优势的路线，那么你可以选择通用的开源模型，比如Qwen2-72B、Llama3-70B，这些模型的能力很强，用来做通用的推理，完全足够。

如果你要做垂直的专业词元，那么你可以选择垂直的开源模型，或者在通用模型的基础上进行微调。比如医疗场景选择Meditron、法律场景选择LawLlama、代码场景选择CodeLlama，这些都是已经做好的垂直模型，可以直接使用。当然你也可以用自己的行业数据对通用模型进行微调，打造属于你自己的专属垂直模型。

这里需要特别提醒的是，你需要注意开源模型的版权协议，避免商用风险。

1．Apache 2.0协议

比如Qwen、Baichuan等模型，都是使用这个协议，允许你免费商用，没有任何限制，不需要申请，也不需要开放你的源代码，是最安全的选择。

2．Llama系列协议

Meta的Llama系列模型做商用的话，如果月收入超过700万美元，则需要向Meta申请授权。对于中小开发者来说，大部分情况下都不会超过这个阈值，但是如果未来规模做大了，就需要注意这个问题。

所以，在选择模型的时候优先选择Apache 2.0协议的模型，这样可以避免未来的版权风险。

9.2 模型的优化：每一种优化，都能把成本降一半

选好了模型之后，接下来就是模型的优化。原生的开源模型推理速度慢、成本高，你需要对它进行优化，才能降低词元成本。

模型的优化有很多种方法，每一种方法都能够帮你把成本降低一大截。

1．量化：用可接受的精度损失换取极致的效率

量化是最常用的模型优化方法，它的原理是把模型的权重从16位的浮点

数，压缩成8位甚至4位的整数。这样模型的大小就会缩小一半甚至更多，计算量也会同步减小，推理速度就会提升一倍以上，而模型的精度损失却非常小。

很多人会担心，量化之后模型的精度会不会下降太多？实际上，根据行业的测试数据，对于绝大多数的场景，量化的精度损失是完全可以接受的。

INT8量化：这是最成熟的量化方案，它把权重压缩到8位整数，模型的大小缩小一半，推理速度提升一倍，而精度损失不到1%，可以忽略不计。对于绝大多数的场景，INT8量化都是最优的选择，平衡了效率和精度。

INT4量化：这是更极致的压缩方案，把权重压缩到4位整数，模型的大小缩小到原来的1/8，推理速度提升2倍以上，精度损失在5%—10%。对于通用的对话场景，这个精度损失是完全可以接受的，但是对于医疗、法律这些对精度要求极高的场景，就需要谨慎使用。

比如，你把一个7B的模型做INT4量化，原来它需要14G的显存，现在只需要3.5G，原来一秒钟能生成1000个词元，现在能生成2000个，你的成本直接降了一半。

现在量化技术已经非常成熟了，GPTQ、AWQ、SmoothQuant这些工具，只需要点几下鼠标，就可以完成模型的量化，不需要你懂专业的技术。很多模型厂商都是用AWQ量化，把模型的成本降低了一半。

2．知识蒸馏：把大模型的能力迁移到小模型

知识蒸馏，就是用一个大的、能力强的“老师模型”来教一个小的“学生模型”。这样小模型就能够拥有大模型90%以上的能力，但是参数却小了很多，推理速度快了很多，成本也低了很多。

比如，你将GPT-4作为老师模型，教一个7B的小模型，最后这个小模型就能够拥有GPT-4 90%的能力，但是推理成本只有GPT-4的5%。这就是知识蒸馏的魔力。

很多中小模型厂商就是用知识蒸馏的方式，把大模型的能力迁移到小模型里，然后用小模型做推理，成本极低，但是能力却很强，能够和大厂

的模型竞争。

现在，LLaMA Factory等工具已经集成了知识蒸馏的功能，不需要自己写代码，只需要上传你的数据，选择老师模型和学生模型，就可以自动完成知识蒸馏，非常简单。

3．推理优化：把硬件的潜力压榨到极致

推理优化就是用vLLM、TensorRT-LLM这些工具，优化模型的推理过程，把硬件的潜力发挥到极致。

原生的transformers推理框架，内存的利用率只有30%。因为KV缓存的碎片化，很多显存都被浪费了，而vLLM用了PagedAttention技术，把KV缓存做成了分页的，就像操作系统的虚拟内存一样，这样内存的利用率就提升到了90%。

所以，原生的transformers推理一秒钟能处理1000个词元，用了vLLM之后能处理5000个，速度提升了4倍，成本直接降低80%。

黄仁勋在演讲中提到的Fireworks AI，就是用了TensorRT-LLM和vLLM，把词元的生成速度从每秒700个，提升到了每秒5000个，成本降了80%。这就是推理优化的力量。

4．**MoE**架构：稀疏激活，降低计算量

MoE，也就是混合专家模型，它把模型分成很多个小的专家模型，每次推理的时候只激活一部分专家，而不是激活整个模型。这样计算量就小了很多、速度快了很多，成本也低了很多。

根据研究报告的统计，MoE模型在同样的精度下，推理成本是dense模型的1/5。也就是说，同样的模型用MoE架构，成本直接降了80%。

DeepSeek的模型就是用了MoE架构，他们的成本只有GPT-4的5%。MiniMax的模型也是用了MoE，所以他们能够把词元价格降到极致。

不过需要注意的是，MoE架构更适合大模型，对于7B、13B这样的小模型，MoE的调度开销会抵消它的优势，所以小模型用稠密架构（dense）反而更合适。

通过这些优化手段，你能够把模型的推理成本大幅降低。同样的模型优化之后，成本只有原来的1/10，这就是你能够比别人更便宜的核心原因。

9.3 微调：**LoRA**技术，把微调成本降低**99%**

如果做垂直领域的业务，那么还需要对模型进行微调，把通用的模型改成你自己的专业模型。

很多人以为，微调大模型要花很多钱、要很多数据，但实际上有了LoRA技术之后，微调的成本已经降到了极致。

LoRA，也就是低秩适应，它的原理是什么？原来的全量微调，你需要微调整个模型的所有参数，比如7B参数的模型，就有70亿个参数，所以你需要很大的算力、很多的数据。

但是LoRA不是这样的，它只在模型的注意力层加了两个小的矩阵，只微调这两个小矩阵的参数，其他的参数都不动，所以它只需要微调0.1%的参数就够了。

所以，你需要的算力就小了99.9%，你需要的数据量也大幅减少。

比如，你要微调一个7B参数的模型，原来的全量微调需要几十张A100，花几万块钱，还要几十万条数据。但是用LoRA，你只需要一张A10，花几千块钱，只要几千条高质量的数据，就可以完成，而且效果和全量微调几乎一样。

这就是为什么LoRA现在这么火，它把微调的门槛降到了极致，让中小开发者也能够定制自己的专属模型。

这就是微调的魔力，可以用很少的钱，很少的数据，就能够把一个通用的开源模型，改成专属的专业模型，建立别人没有的竞争壁垒。

9.4 真实案例：**MiniMax**，如何靠模型迭代把成本降低**90%**

MiniMax之所以能够把词元价格做到大厂的1/17，靠的就是持续的模型优化。

刚开始的时候，MiniMax也是用的普通模型，成本很高。2024年，他们的M1模型，成本是每百万词元1美元，日活跃用户数只有100万。于是他们开始持续的模型优化。

首先，他们用了MoE的架构，做了稀疏激活，把计算量降了下来。原来每次推理要激活整个模型，现在只要激活1/5的专家，计算量直接降了80%。

之后，他们推出了M2模型，成本降到了每百万词元0.5美元，价格定为1美元，瞬间涨到了1000万日活。

然后，他们又做了量化优化，用了AWQ量化，把模型的权重压缩成4位，速度又提了一倍，成本又降了一半。他们用了vLLM和TensorRT-LLM，做推理优化，把推理的速度又提了好几倍，成本又降了很多。

他们推出了M2.5模型，成本降到了每百万词元0.3美元，价格定为0.5美元，涨到了5000万日活。

随后，他们又使用了新的Blackwell架构，加上新的软件优化，成本又继续下降，降到了每百万词元0.1美元，价格定为0.3美元，涨到了1亿日活。

最后，他们把单位词元的成本，降到了GPT成本的5%，然后他们把价格定到了每百万词元0.3美元，把所有的竞争对手都打趴下了。

接下来，他们的用户开始爆发式增长，规模越来越大，成本越来越低。他们又把价格降了一点，吸引更多的用户，形成了正向的飞轮。

现在，MiniMax在港股的市值已经超过了3000亿港元，他们只用了几年的时间，就做到了这个成绩。这个案例告诉我们，模型的优化是词元生产的核心，谁能够把模型的成本降下来，谁就能够占领整个市场。

第十章 部署：如何把模型变成可售卖的词元**API**

搞定了算力，搞定了模型，接下来就是部署的环节。要把模型变成一个可调用的API，然后加上认证、计费的功能，让用户能够调用词元，支付费用。

10.1 一键部署：把模型变成兼容的**API**服务

很多人以为，部署大模型是很复杂的事情，要写很多代码，要懂很多网络知识。但实际上，现在已经有很多成熟的一键部署工具，只需要运行一行命令，就可以把模型部署成可调用的API。

最常用的工具就是VLLM，还有Text Generation Inference，以及Ollama。

比如，使用VLLM，只需要运行以下命令：

python -m vLLM.entrypoints.openai.api_server --model your_model_path --quantization awq --host 0.0.0.0 --port 8000 --api-key your_secure_key

运行这行命令之后，模型就具备了与OpenAI完全兼容的API接口。用户调用你的API，就和调用OpenAI的API完全一样，不需要修改任何代码。他们原来用OpenAI的代码，只要把地址改成你的，就可以直接使用你的词元了。

这里我们来解释一下每个参数的含义，方便根据自己的情况调整：

--model：指定你的模型文件的路径，也就是你微调之后的模型的位置。

--quantization：指定量化的方式，如果你用了AWQ量化，就填awq，如果是GPTQ就填gptq，这样VLLM就会加载对应的量化优化。

--tensor-parallel-size：张量并行的数量，如果你的模型很大，比如70B的，需要多张显卡才能跑，你就用这个参数来设置并行的显卡数量，比如你有2张卡，就填2。

--host：绑定的IP地址，0.0.0.0代表允许所有外部地址访问，如果只是本

地测试，就填127.0.0.1。

--port：绑定的端口，可以自己指定，只要没有被占用即可。

--api-key：设置API的密钥，这样只有带了这个密钥的请求才能访问你的API，防止别人盗用你的接口。

就这么简单，你不需要懂什么网络知识，不需要懂什么并发处理，只要运行这一行命令，就搞定了部署。如果你不想自己搭建，也可以用Hugging Face的推理端点，或者硅基流动的部署服务，把模型上传上去，它们帮你部署，你只要付费即可，哪怕完全不懂技术也能搞定。

10.2 计费与认证：让用户为你的调用付费

部署好API之后，接下来就是计费的功能。你要统计每个用户调用了多少词元，然后收取对应的费用。

这个也非常简单，因为VLLM的API会自动返回每个请求的词元数量，你只需要在网关统计每个用户的用量，然后扣除对应的余额即可。

你可以用一些开源的网关，如Kong或APISIX，这些网关已经帮你做好了认证、限流、统计的功能，你只要简单配置一下，就可以使用了。

比如，用户注册的时候，你给他生成一个API Key，用户调用你的API的时候，网关会验证他的Key，然后统计他的词元用量，扣除他的余额。如果他的余额不够，就拒绝他的请求。

就这么简单，你不需要自己写复杂的计费系统，这些开源工具已经帮你做好了。如果觉得自己搭建麻烦，也可以用第三方的计费服务，如Stripe或Paddle Billing，它们帮你做支付和计费，你只要对接一下即可。国内的话，可以用微信支付和支付宝，对接起来也非常简单。

10.3 对接模型市场：让全球的用户都能买到你的词元

部署好了API之后，接下来就是寻找用户。你不需要自己去做推广，可以把你的模型放到模型市场上，它们会给你带来海量的用户。

最有名的就是OpenRouter，这是一个模型聚合平台，上面有几百万的开

发者，他们在平台上找模型、买词元。只要把你的API接到OpenRouter上，你的模型就会出现在他们的平台上，几百万的开发者都能看到你的模型，调用你的词元，给你付费。

对接OpenRouter的步骤也非常简单，最新流程只需要5步：

1．注册**OpenRouter**账号

访问OpenRouter的官网，用邮箱注册账号，只需要邮箱验证，不需要身份证、手机号等信息，5分钟就能完成。

2．进入提供者后台

注册完成之后，进入提供者后台，选择添加新的模型。

3．填写模型信息

填写模型的名字、描述、分类，还有你的API端点，设置你的定价。

4．等待审核

提交之后，平台会对你的模型进行审核，主要检查可用性和稳定性，一般1-2天审核就会通过。

5．开始变现

审核通过之后，你的模型就会上线，用户就可以调用你的模型，你就能获得收益了。

MiniMax就是这么做的，他们把自己的模型放到了OpenRouter上，一下子就获得了全球的用户，他们的调用量瞬间爆发，直接登顶全球第一。

国内有硅基流动、魔搭社区，这些都是模型市场，你把你的模型放上去，就有国内的用户来买你的词元。

这太方便了，你不需要自己做产品，不需要自己做推广，你只要把你的模型放上去，就有用户来买，你等着收钱就行了。对于第一次做词元生产线的你来说，这是最快的获客方式。

10.4 高并发与自动扩缩容：应对用户的爆发增长

当你的用户越来越多，调用量越来越大，你怎么应对高并发的流量？怎么保证你的服务不会挂掉？

这个也很简单，你可以用K8s做自动扩缩容。也就是说，当用户多的时候，自动增加服务器和算力；当用户少的时候，自动减少服务器和算力。这样，你就能应对突发的流量，而且不会浪费你的算力成本。

比如，白天你的调用量是10000QPS，就自动启动10台服务器。晚上，调用量降到1000，就自动关掉9台，只留1台，节省电费。

具体的配置也很简单，你可以用K8s的
HPA（Horizontal Pod Autoscaler），根据GPU的利用率来自动扩缩容。比如，
apiVersion: autoscaling/v2kind: HorizontalPodAutoscalermet adata: name: vll
hpaspec: scaleTargetRef: apiVersion: apps/v1 kind: Deployment name: vllm-deployment minReplicas: 1 maxReplicas: 100 metrics: - type: Resource resou

这个配置的意思是，当GPU的平均利用率超过70%的时候，就自动增加节点，当低于30%的时候，就自动减少节点，最小保留1个节点，最多可以扩到100个节点。这样，不管流量怎么涨，你的系统都能自动应对。

很多大模型厂商都是这么做的，如Fireworks AI，他们的自动扩缩容系统，能在几秒钟内把算力从0台扩至1000台，以应对突发的流量，而且成本极低。

除此之外，你还可以做负载均衡，把用户的请求分散到不同的服务器上，保证服务的稳定。

这样，无论你的用户有多少，都能应对。哪怕你的用户一天涨了10倍，也不用担心，你的系统会自动扩缩容，帮你搞定。

10.5 部署的安全防护：避免你的**API**被滥用

很多开发者在部署的时候会忽略安全问题，导致API被别人盗用，或者被刷量，产生很高的成本。这里我们介绍几个基础的安全防护措施。

1．启用**API**密钥

就是我们之前提到的，在VLLM里设置--api-key，只有携带了正确密钥的请求才能访问，防止匿名用户盗用你的接口。

2．配置跨域限制

用--allowed-origins参数，只允许你指定的域名跨域访问，防止其他网站盗用你的接口。

3．禁用接口文档

用--disable-fastapi-docs参数，禁用Swagger UI的接口文档，防止你的接口被暴露。

4．启用**HTTPS**

用--ssl-certfile和--ssl-keyfile参数，启用HTTPS加密，防止请求被劫持。

5．**DDoS**防护

可以用Cloudflare或者云厂商的DDoS防护服务，防止被恶意的流量攻击，避免你的服务瘫痪。

这些简单的防护措施，能够避免绝大多数的安全问题，防止你的API被滥用，减少不必要的成本损失。

10.6 边缘部署：把词元生产放到距用户最近的地方

如果你是面向企业用户的，比如工厂、医院，这些对数据隐私要求很高的用户，你还可以做边缘部署。就是把你的模型部署到用户的设备上，或者边缘的节点上，这样用户不需要调用云端的API，直接在本地就能推理，成本更低，延迟更低，而且数据不需要出用户的机房，安全和隐私都能保证。

开鸿智谷把模型做成了AI BOX，部署到工厂的边缘、发电厂的旁边。用户在本地就能推理，不需要消耗云端的算力，成本直接降到了最低。

比如，有一家钢铁企业，原来使用云端的词元，每百万词元需支付10元

钱，而且数据要传到云端，不但有隐私风险，而且还会延迟100ms。使用边缘部署之后，每百万词元只需支付1元钱，成本降了90%，数据不需要出工厂，完全满足隐私要求，延迟也降至10ms，满足了工业生产的实时性要求。

其实对于企业用户来说，他们更愿意为边缘部署支付更高的费用，因为这能解决他们的隐私问题和延迟问题。很多企业宁愿花10倍的钱，也要使用边缘部署的词元，因为安全对他们来说太重要了。

所以如果你是面向企业用户的，边缘部署是一个非常好的选择，能帮助你获得更高的利润。

第十一章 运营：如何让你的生产线自动赚钱

当词元生产线搭建并部署完毕，接下来就是运营的环节。要让生产线自动赚钱，让财富自动增长，让飞轮转起来。

11.1 定价：用合理的价格，启动你的增长飞轮

首先，就是定价的问题，这取决于你走的是哪条财富路径。

1．成本路线：用性价比击穿市场

如果你是做通用词元，走成本优势的路线，那你就要用极致的性价比，击穿市场，启动你的增长飞轮。

行业的通用策略是，定价为成本的3倍。

比如，每百万词元成本是0.1美元，那定价就为0.3美元，比大厂的5美元低很多。这样开发者就会来使用你的词元，用户就会越来越多，规模也会越来越大，成本就会越来越低。然后，就可以把价格再降一点，吸引更多的用户，这样飞轮就转起来了。

MiniMax就是这么做的，他们的成本降到了0.1美元，然后他们定0.3美元，比所有的竞争对手定价都低，然后用户爆发，规模扩大，成本再降，然后价格再降。就是这么简单的逻辑，最后垄断了市场。

很多人会问，定价这么低，能赚钱吗？当然能。因为你的成本只有0.1美元，你定0.3美元，还有0.2美元的利润。而且用户越多、规模越大，成本就越低，利润就越高。

2．价值路线：按用户的价值定价

如果你是做垂直词元，走价值优势的路线，那定价就可以高一点。因为你的词元价值高，企业愿意付费。

行业的通用策略是，定价为用户ROI的10%。也就是说，你的词元能帮用户赚100元，你收费10元，用户完全能够接受，因为他还有90元的利润。

比如，创研股份的营销词元，他们的词元能帮助营销公司提高300%的效率，一家营销公司如果用了他们的词元，每个月能多赚10万元，那么支付1万元的词元费用，营销公司完全愿意。因为这些钱对他们来说，只是利润的1/10。

他们的词元定价是每百万词元1000元，是通用产品的2000倍，但是营销公司还是抢着买，因为价值在那里。

所以，定价要根据路径来定：成本路线，就打性价比，启动飞轮；价值路线，就按价值定价，收割利润。

11.2 用户运营：让用户持续用，持续付钱

然后，就是用户运营。要让用户持续用你的词元，持续给你付费。

1．通用开发者用户：做高留存的生态

对于通用的开发者用户，你要做的，就是把词元做得更稳定、更便宜。不断优化模型，优化部署和成本。

你还可以做开发者社区，在Discord或QQ群里与用户互动，收集他们的反馈。比如，用户觉得模型哪里不好，你就优化哪里。你还可以做详细的文档，做示例代码，帮助用户快速接入，降低他们的使用门槛。

除此之外，还可以做词元包。比如，1000万词元卖2元钱，用户买了词元包就会持续使用，不会流失。还可以做订阅，如每个月10元钱随便用，这样用户的复购率就会很高。

2．垂直企业用户：做客户成功

对于垂直的企业用户，你要做的，就是给他们提供优质服务，帮他们解决问题。比如，给企业用户做工业词元，你要帮他们落地，帮他们解决生产中的问题，让他们看到价值，他们就会持续用你的词元，持续给你付费。

你可以做专属的客户经理，帮用户对接，帮他们解决问题，定期回访，看他们的使用情况，帮他们优化。你还可以做定制化的服务，给企业做专属的微调和部署，收取服务费，这样你的收入就更多了。很多做垂直

词元的厂商，他们的服务费收入比卖词元的收入还要高。

11.3 数据合规：合法利用用户数据优化模型

当用户使用你的词元，你就会得到很多数据，你可以利用这些数据不断优化模型，让词元越来越好。

但这里有一个非常重要的问题：数据的合规性。很多人会直接收集用户的提问和回答用来训练模型，但如果不合规，会违反个人信息保护法，带来法律风险。

可以通过以下几种方式，来保证合规。

1．在用户协议中告知

在用户协议里明确告知用户，我们会收集匿名化后的使用数据用来优化模型，并且用户可以选择关闭数据收集。

2．数据匿名化

收集数据的时候，要去掉所有的个人信息，如用户的API Key、身份信息，还有提问里的个人信息，把数据完全匿名化，这样就不会涉及个人信息泄露的问题。

3．数据本地化

如果用户在欧洲，要符合GDPR的要求，把数据存储在欧洲的节点，给用户提供数据删除的权利。

通过这些措施，你就可以合法地收集和使用用户数据，用来优化模型，不会有不合规的风险。

例如，用户调用模型，你把匿名化后的提问和回答收集起来，然后利用这些数据继续微调模型，让模型越来越懂用户，越来越准。这样你的词元质量就会越来越高，用户就会越来越多，壁垒就会越来越高。这就是数据飞轮的力量：用户越多，数据越多，模型越好，用户越多。这就是“护城河”的作用，用户带来的数据是别人拿不走的，是你独有的壁垒。

11.4 飞轮法则：让你的财富自动增长，自动复利

当你把这些都做好了，你的飞轮就转起来了。

第一个飞轮：用户越多→规模越大→成本越低→价格越低→用户越多。

第二个飞轮：用户越多→数据越多→模型越好→词元质量越高→用户越多。

这两个飞轮一起转，你的财富就会自动增长，自动复利。

你不需要做什么，只要把生产线搭好，飞轮就会自己转，用户会自己来，钱会自己进账。黄仁勋在演讲里提到了CUDA的飞轮。他说，CUDA用了20年时间从一个小的架构变成了万亿的生态，靠的就是这个飞轮。词元经济的飞轮比CUDA的飞轮转得更快，因为词元的边际成本更低，规模效应更强。

11.5 生产线的维护：让你的印钞机持续运转

很多人以为，搭建了生产线之后就可以完全不管了，就可以躺着赚钱。但实际上，被动收入不是完全不用管，还是需要做一点简单维护，让生产线能够持续运转。

主要的维护工作有以下两个。

1．定期更新模型

新的开源模型会不断被开发出来，如Qwen2.5、Llama3.1，它们性能更强、成本更低，你需要定期把模型更新到最新的基础模型，重新微调，这样才能保持竞争力，不会被淘汰。

2．日常监控

你需要监控服务的状态，如可用性、延迟、错误率、是否有故障，每天看一下，出了问题及时处理，保证服务的稳定。

这些工作非常简单，每天只需要1小时左右的时间就可以完成。剩下的时间，你完全可以自由安排，去旅游、去陪伴家人、去做你想做的事，

你的收入不会停。

11.6 被动收入：拥有你的“印钞机”，享受复利

词元生产线能带来持续的、指数级增长的被动收入，它能让人真正地实现财富自由。很多人搭建了自己的词元生产线之后就实现了财富自由。他们不再需要为了钱而工作，他们的资产在为他们工作，这就是为什么说词元生产线是AI时代的“印钞机”。

现在词元经济的浪潮才刚刚开始。根据研究报告的数据显示，到2030年，我国词元经济的产业规模会达到3万亿人民币。这是一个巨大的市场，充满了无数的机会，而现在就是最好的时机。动手去做，搭建你的第一条词元生产线，抓住这个时代的红利，实现你的财富自由。

第四部分

词元商业化：把生产线变成印钞机

在词元经济的框架下，尽管核心都是“按词元计费”，但不同的市场参与者根据自身的资源禀赋和战略定位，演化出了几种主流的商业模式。

商业模式	核心逻辑	代表性企业	优势	劣势
基础设施即服务（IaaS）	销售“生产词元的能力”，核心产品是算力（GPU集群），客户自行负责模型部署和管理	英伟达（芯片销售）、光环新网（算力租赁）	利润率高（销售实体硬件），不直接参与AI服务的复杂运营和定价	客户群体专业门槛高，受硬件供应链和制造成本影响大
平台即服务（PaaS）/模型即服务（MaaS）	销售“精炼好的词元”，提供集成了算力、模型、开发工具的平台，客户通过API调用服务	阿里云、腾讯云、AWS、OpenAI	生态优势明显，客户群体广泛；商业化模式成熟，收入稳定；降低客户使用AI的技术门槛	竞争激烈，利润空间受挤压；需要持续投入巨资进行模型研发和基础设施升级

续表

商业模式	核心逻辑	代表性企业	优势	劣势
垂直行业应用（SaaS）	销售“解决特定问题的词元包”，将AI能力封装在针对特定行业（如金融、医疗、法律）的软件中	科大讯飞（星火大模型在教育/医疗的应用）	价值明确，客户付费意愿高；产品黏性强，不易被替代	市场规模受限于特定行业，需要深厚的行业知识
聚合与分发平台	连接“词元生产者”和“消费者”，整合多家模型和算力供应商的API，提供统一的服务入口和定价	OpenRouter（海外）	为客户提供更多选择和价格透明度，帮助中小模型厂商快速触达市场	自身利润率较低，价值相对轻，容易被巨头绕过

当你搭建好了自己的词元生产线，拥有了稳定的词元生产能力，接下来的核心问题，就是如何对接这些成熟的商业模式，将这套生产能力转化为持续、稳定、可增长的现金流，把生产线真正变成一台自动运转的“印钞机”。

词元经济的商业化与传统的AI创业有着本质的区别。传统的AI创业，需要投入大量的资金做产品、做运营、做推广，边际成本高、增长慢、盈

利难。而词元商业化，是把标准化的词元作为产品，依托已经成熟的模型市场、开发者生态、全球分发网络，实现低成本获客、规模化变现、复利式增长。

在这一部分中，我们将系统拆解词元商业化的完整路径，从词元工厂变现模式，到低成本规模化的增长逻辑，再到高价值垂直领域的利润挖掘，以及词元出海的全球化机遇，最后到智能体时代的增量爆发，完整呈现如何把生产线变成一台持续生财的“印钞机”。

第十二章 词元工厂变现模式

在词元经济的商业版图中，上游的词元生产者处于整个价值链条的顶端，掌握着词元的生产、供给、定价与分发核心权力，是AI时代真正的“能源供应商”。与下游消耗词元的工具使用者、中游倒卖词元的服务商不同，上游玩家的核心竞争力，是建立可持续、可规模化、可复利的词元供给体系，将算力、数据、电力转化为源源不断的现金流。

从产业实践来看，经过2024—2026年的市场验证，全球词元经济上游已经形成五大商业化变现模式，分别是：词元引擎优化（Token Engine Optimization, TEO）、垂直词元订阅、企业级私有词元服务、词元出海服务、算力租赁+词元打包服务。这五大模式覆盖了个人、中小团队、中型企业、大型集团、跨境机构等全类型玩家，每一种模式都有清晰的盈利逻辑、成本结构、客户群体与增长路径，共同构成了词元工厂商业化变现的底层框架。

12.1 词元引擎优化：面向开发者的效率革命

词元引擎优化，是上游词元玩家最基础、最普惠的商业模式，核心是为开发者、中小模型厂商、AI应用服务商提供更低成本、更高效率、更稳定的词元推理服务，通过技术优化降低单位词元成本，提升词元生成速度与质量，赚取效率差价。

在AI应用爆发的背景下，全球开发者群体呈现指数级增长。根据2026年全球开发者联盟数据显示，全球AI开发者数量已突破3000万，其中90%为个人开发者、小型创业团队与中小SaaS厂商，这类群体没有能力自建算力集群与模型生产线，却有刚性的词元调用需求。但传统大厂API价格高昂（如GPT-4每百万词元收费5美元以上）、响应不稳定、对中文与垂直场景支持不足，成为制约中小开发者落地AI应用的核心瓶颈。

TEO模式的核心价值，正是解决这一痛点。其商业逻辑可以概括为：低成本生产词元→标准化封装API→对接海量中小客户→按用量计费→通过规模效应持续降本。与传统模型厂商不同，TEO玩家不追求打造全能大模型，而是聚焦推理效率、成本控制和服务稳定性，把生产便宜词元这件事做到极致。

从技术架构来看，TEO模式的核心支撑包括以下四大模块。

1．高效推理引擎

基于VLLM、TensorRT-LLM、Text Generation Inference等开源工具，结合模型量化、知识蒸馏、MoE稀疏激活技术，将推理速度提升5～10倍，单位词元成本降低80%以上。例如，黄仁勋提到的Fireworks AI案例，在不更换任何硬件的前提下，仅通过软件栈与推理算法优化，词元生成速度从每秒700个提升至近5000个，成本直接下降85%。

2．智能算力调度

整合西部低价算力、云厂商竞价实例、全球闲置GPU资源，通过动态调度系统将用户请求分发至成本最低的算力节点，实现“同一服务、不同成本、最优收益”。例如，CoreWeave通过整合全球闲置GPU，算力成本比传统云厂商低30%-50%，为TEO模式提供了底层支撑。

3．标准化**API**接口

兼容OpenAI API协议，开发者无须修改代码，仅更换接口地址即可调用，降低接入成本。这一设计是TEO模式快速获客的关键，也是OpenRouter、硅基流动等聚合平台快速崛起的核心原因。

4．轻量化运营体系

无须庞大的销售与客服团队，通过模型市场、开发者社区、开源平台实现自然获客，运营成本极低。个人开发者或3-5人小团队即可完成全流程运营，边际成本趋近于零。

从盈利模式来看，TEO玩家采用“按量计费+阶梯定价”的双重策略：基础定价远低于大厂（通常为大厂价格的1/5～1/3），针对高用量客户提供阶梯折扣，用量越大单价越低，以此锁定长期用户。

从行业数据来看，TEO模式是增长最快的词元商业模式。以OpenRouter平台为例，其已形成年化数千万至过亿美元营收、月处理数十万亿词元、服务数百万开发者的成熟服务生态，并持续以10倍的年增长率扩张，是典型的“轻资产、高利润、快增长”模式。

12.2 垂直词元订阅：行业专属的长期收益

垂直词元订阅，是上游词元玩家高利润、高壁垒的商业模式，核心是针对医疗、法律、金融、制造、教育等高付费行业，打造专属专业词元，以包年/包月订阅的方式，为企业提供稳定、合规、高精度的词元服务，赚取价值溢价。

与TEO模式的通用词元、低价走量方法不同，垂直词元订阅聚焦专业词元、高价稳收，其核心逻辑是：行业知识沉淀→专用模型微调→高质量词元生产→企业订阅付费→数据反哺优化→形成行业壁垒。在词元经济体系中，通用词元是基础能源，而垂直词元是高端燃料，企业愿意为垂直词元支付10～100倍的溢价，因为专业词元能够直接解决生产经营中的核心痛点，创造可量化的商业价值。

根据研究报告数据显示，垂直领域词元的单位价值是通用词元的20～100倍，毛利率可达90%以上，且客户续约率超过85%，是词元商业化中最优质的商业模式。目前，垂直词元订阅已在营销、工业、医疗、法律、服装等领域落地成熟应用，验证了其商业可行性。

垂直词元订阅的客户群体，主要是高付费能力的行业企业，这类客户对价格不敏感，更关注词元的准确性、合规性、安全性，只要你的词元能够帮他们解决核心问题，他们愿意支付长期的订阅费用，这就为词元生产者带来了稳定、持续的现金流。

12.3 企业级私有词元服务：安全可控的定制化供给

企业级私有词元服务，是针对大型企业、政府机构、金融机构等对数据隐私要求极高的客户提供的内网部署、安全隔离、合规可控的专属词元服务，是词元商业化中客单价最高的商业模式。

对于大型企业来说，通用的公有云词元服务，存在泄露数据隐私的风险：企业的核心数据、业务数据、客户数据，不能传到公有云的模型里，否则会有泄露的风险。例如，金融机构的客户数据、制造业的生产数据、政府的政务数据，这些都是核心机密，不能出企业的内网。

所以，这类客户需要私有部署的词元服务：把模型部署到企业的内网里，数据不需要出企业的机房，所有的推理都在本地完成，既满足了企业的AI需求，又保证了数据的安全和隐私。

这种模式的客单价极高，通常一个企业的私有词元服务，年费就可以达到几十万甚至上百万，毛利率也极高，因为模型已经做好了，部署成本很低，后续的维护成本也很低。

除此之外，私有词元服务还可以做定制化的微调，根据企业的专属数据，微调专属的模型，让模型更适配企业的业务场景，进一步提升价值。例如，某银行用自己的风控数据，微调了金融专属模型，部署到内网，做风险评估，效果比通用模型好很多，而且数据完全安全。

12.4 词元出海服务：中国算力的全球化变现

词元出海服务，是把中国的低成本词元，卖给全球的用户，赚取全球化的溢价，这是中国词元从业者独有的优势，因为中国有全球最便宜的绿电、最便宜的算力、最高效的模型，我们的词元成本，只有海外的1/10～1/6，在全球市场有极强的竞争力。

根据研究报告数据显示，2026年全球词元市场的规模已经突破8000亿美元，海外的词元价格比国内词元价格高很多，如OpenAI的GPT-4，每百万词元要5美元；Anthropic的Claude Opus，每百万词元要25美元。而中国的MiniMax的M2.5模型，每百万词元只要0.3美元，价格只有它们的1/17，但是能力却接近GPT-4，这样的性价比，在全球市场是碾压级的。

词元出海的模式非常简单，既不需要去海外做推广，又不需要建海外的机房，只要把你的API接到OpenRouter这样的全球模型市场，他们就会帮你把模型API卖给全球的用户，并帮你做结算和交付，你只要负责生产词元就行了。

这种模式，让中国的中小词元生产者也能轻松做全球生意。不需要懂英语，不需要做海外推广，靠OpenRouter的平台，就能赚到全球的钱。

12.5 算力租赁+词元打包：一站式的企业解决方案

算力租赁+词元打包，是针对有算力资源的玩家，把算力租赁和词元服务打包卖给企业客户，一站式解决企业的AI需求，这是一种重资产但高壁垒的商业模式。

很多企业想要做AI，但是他们不懂算力、不懂模型、不懂词元生产，他

们不知道怎么把算力变成词元，所以他们需要一站式的解决方案：有人提供算力、模型和词元服务，而企业只需付费即可。

CoreWeave就是这个模式的开创者，他们整合了全球的闲置GPU，然后把算力和词元服务打包，卖给AI公司，他们的价格比AWS低30%-50%，但是利润却很高，因为他们的算力成本很低。现在，CoreWeave的年收入已经超过20亿美元，企业估值超过200亿美元，成为AI原生云的龙头。

国内也有很多这样的玩家，比如光环新网，他们在呼和浩特的智算中心有大量的低价绿电算力，他们把算力和词元服务打包，卖给企业客户，企业只要付费就能直接使用AI服务，不需要付出额外的管理成本，非常方便。

这种模式的壁垒很高，因为词元生产者需要整合大量的算力资源、需要搭建完整的词元生产体系、需要强大的销售和服务能力，中小玩家很难进入，但是一旦做成了，就会有很强的规模效应，成本会越来越低，壁垒会越来越高。

第十三章 低成本词元：规模化横扫市场

在词元经济的整套商业逻辑里，低成本词元是一条完全靠效率取胜、靠规模站稳脚跟、靠成本优势筑起“护城河”的主流路线。走这条路，不用纠结单个词元能卖出多高的价格，也不用深耕某个垂直行业建立专业壁垒，核心思路只有一条：把整条生产链条的成本做到极致，用普通人、小企业都无法拒绝的低价，快速拿下海量用户，再在用户不断增长的过程中把成本压得更低，使竞争力更强，最终形成越做越大、越做越稳的正向循环，彻底占领整个通用词元市场。

对绝大多数普通人、中小团队和创业公司来说，做低成本词元不是薄利多销的辛苦生意，而是最容易上手、最容易做出规模、最容易实现长期稳定收益的模式。

从行业发展的规律来看，做低成本词元的模式，融合了互联网的低价获客思维、工业时代的规模效应，以及AI时代的算力革命优势。传统互联网靠低价甚至免费吸引用户，再靠流量变现；传统工厂靠扩大产量降低单件成本；而做低成本词元的模式可以把两者的优势合二为一，先用极具优势的低价快速吸引大量对价格敏感的用户，等用户规模上来之后，再靠规模把生产成本持续降低，最后靠用户的使用习惯和生态绑定留住长期客户，一步步从价格优势变成规模优势，再从规模优势变成别人抢不走的竞争壁垒。

在黄仁勋提出的“每瓦电产出多少词元”这一核心指标下，低成本路线的价值被放到最大，每一次效率提升、每一次规模扩大，都会直接变成更强的市场竞争力和更可观的利润，成为横扫市场的关键。

13.1 价格策略：用极致性价比击穿市场

低价，是低成本词元模式打开市场、吸引第一批用户、淘汰竞争对手最直接的武器。在通用词元市场上，各家提供的服务质量差距并不大，用户不会为了品牌或者花哨的技术多花钱，真正决定他们选择的，就是稳定、好用、够便宜。当不同商家的词元在生成效果、响应速度、接口适配方面都差不多时，价格就成了唯一的决定因素。做低成本词元，定价不是盲目赔本赚吆喝，而是依托从头到尾的成本控制能力，把价格定到大厂价格的30%-70%，在保证自己有钱赚的前提下，直接打破市场原有

的价格规则，让用户主动放弃高价服务，转而来用你的词元。

之所以能把价格压得这么低，其根本原因是巨大的成本差距。像OpenAI、Anthropic这样的海外大厂，还有国内的头部科技公司，他们的词元生产都依赖自建的高端数据中心、自研的大模型、标准化的运维团队，不仅要承担巨额研发和管理费用，而且数据中心大多建在电费昂贵的大城市，电力成本一直居高不下。各种成本加在一起，他们的词元生产成本根本降不下来，定价也必须维持在高位才能保证利润，根本无法进入低价市场。

而你做低成本词元，成本优化是从头到尾、覆盖每一个环节的。在用电上，直接对接国家“东数西算”的西部枢纽节点，用风电、光伏这类绿色电力，每千瓦时电量只要0.15～0.3元，比大城市便宜一半还多，从源头把最大的成本压了下来；在算力上，不用追求昂贵的专用芯片，改用性价比更高的通用推理芯片，再租用云厂商的闲置算力资源，通过批量采购和长期合作降低硬件成本；在模型上，不用从零开始训练超大模型，直接用成熟的开源模型做轻量化优化，把算力消耗大幅降低；在运营上，用智能调度和推理优化技术，让显卡的利用率大幅提升，避免算力闲置浪费。

这4个环节的优化叠加在一起，使词元生产成本可以降至头部大厂的1/5甚至1/10。哪怕把价格定到他们的三成，依然可以有一半以上的利润，形成“价格更低、利润更高”的优势。在这种优势面前，价格战不再是恶性竞争，而是自然淘汰落后玩家的“过滤器”。

市面上那些没有技术优化能力，只是转手倒卖大厂API的中小商家，成本和大厂几乎一样，根本不敢跟着降价，一降价就亏损，只能眼睁睁看着用户流失，最后退出市场。而你则可以进入“降价—引来更多用户—成本进一步降低—继续降价”的良性循环，慢慢把大厂的通用词元市场份额抢过来，快速站稳脚跟。

市场上已经有非常成功的例子，MiniMax和DeepSeek就是靠这条路崛起的。MiniMax依托西部低价绿电和极致的效率优化，把输入词元价格定到每百万词元0.3美元，只有Claude Opus的1/17，服务质量却相差不大，再加上接口完全兼容，很快就成为全球调用量最高的通用词元服务；DeepSeek通过模型压缩和分布式算力，把价格压到更低，迅速占领了中小开发者市场。这足以证明，在通用词元市场，性价比就是最强的获客能力，只要成本优势足够大，便可通过低价轻松撬动海量用户。

13.2 客户定位：抓住存量市场的刚性需求

做低成本词元，要抓住的核心客户，是数量极其庞大、对价格高度敏感、需求又非常刚性的群体。这部分用户包括中小AI开发者、独立技术团队、中小SaaS公司、各类工具开发商、AI镜像网站运营者等。他们有几个共同特点：必须高频使用词元，自身盈利空间有限，非常在意API的成本，只要稳定兼容、价格更低，就愿意更换服务商，而且他们早就被大厂的高价压得喘不过气，一直盼着更便宜的替代方案。

在低成本词元出现之前，这些用户别无选择，只能用大厂的高价API，长期处于成本增长的困境里。中小开发者做聊天机器人、文案工具、代码助手，一半的收入可能都要花在词元调用上，很多好项目因为成本太高根本赚不到钱；中小型SaaS公司想在产品里加入AI功能提升竞争力，可词元成本太高，要么不敢增加功能，要么定价太高卖不动；AI镜像网站更是完全靠词元差价赚钱，大厂价格太高，他们几乎无利可图，随时可能亏本。

这些用户的需求有3个关键优势：一是刚性，AI功能已经是他们产品的核心，不可能停用；二是高频，每天都有稳定的请求量，词元消耗持续稳定；三是迁移成本极低，目前主流词元服务都兼容OpenAI的接口规范，用户不用改代码、不用调流程、不用额外花钱，只换一个接口地址就能完成切换。一边是高昂的成本压力，一边是几乎为零的切换成本，你的低价词元一出现，对用户来说就是最优解，根本不需要花费大力气推广，用户就会主动找上门。

对你来说，低成本词元策略不只是低价卖货，更是在帮这些中小主体降低成本、提高利润。开发者成本降90%，利润就能大幅增长，原本不赚钱的项目也能正常盈利；SaaS公司成本降80%，就可以下调产品价格，快速吸引更多用户，收入反而增长；镜像网站有了低价词元，差价空间变大，生意也能稳定做下去。这种双赢的关系，会让用户对你的词元产生很强的依赖性，他们的业务越做越大，词元消耗量就越来越多，你的规模也会跟着水涨船高。

这个存量市场的规模，远比大多数人想象的要大。截至2026年，全球中小型AI开发者已经超过3000万人，中小型SaaS厂商超过2000万家，各类工具开发商也有500万家以上，他们消耗的词元占整个通用词元市场的七成以上。头部大厂只顾着服务高利润的大客户，长期忽略了这个庞大的市场，留下了巨大的空白。只要抓住这群用户，不用和大厂争抢高端

客户，就能靠海量的中小用户把规模做起来，这正是低成本路线最核心的优势。

13.3 规模飞轮：放量之后，成本再降，壁垒再高

当依靠极致低价吸引了大量用户，词元出货量真正起来之后，一套自动放大的规模飞轮就会启动。这套飞轮由4个部分共同推动：算力议价能力变强、模型推理效率更高、调度系统更智能、用户生态更牢固，形成一个别人无法复制、无法追赶的循环。简单来说，用户越多，你的成本越低；成本越低，你能吸引的用户就更多，循环不断加速，最终形成别人跨不过去的壁垒。

第一个加速点，是算力采购的议价能力变强。词元生产中，算力租赁和折旧是一大笔开销，而采购量直接决定单价。刚开始用量小，只能按标准价格租赁算力；等每天词元出货量从几百万变成几十亿，你就成了云厂商的核心客户，可以签订长期合同，拿到30%-50%的折扣，还能优先拿到紧缺的显卡资源。规模每扩大1倍，算力单价就能再降一档，成本优势越来越明显。

第二个加速点，是推理效率持续提高。海量的用户请求和真实的使用数据，是优化模型最好的原料。你可以根据用户的实际使用习惯，优化批处理、缓存、模型压缩等技术，让每一度电、每一张显卡产出更多词元。例如，根据用户常用请求优化缓存，能减少三成的重复计算；用海量数据微调模型，速度能再提升1倍。用户规模越大、数据越多，优化效果越好，单位成本就越低。

第三个加速点，是智能调度把闲置算力利用殆尽。当你有了大量用户和多个算力节点，就可以把请求自动分配到电费最便宜、算力最空闲的节点，不浪费一点资源。大规模的请求还能削峰填谷，避免算力闲置，让显卡利用率从70%提升至95%以上。同样的算力设备，能处理更多的词元请求，固定成本被进一步摊薄。

第四个加速点，是生态绑定形成真正的壁垒。当大量中小型开发者基于你的词元接口开发应用、搭建系统、跑通业务之后，他们的整个业务流程都与你的词元服务深度绑定，如果要换成其他词元服务，需要改代码、改系统、重新适配，成本高到无法接受。这时候，就算别人降价，也挖不走你的用户。同时，你的客户越多，愿意围绕你的生态做开发的开发者就越多，你的服务就会变得更完善、更好用，形成“用户越多、

生态越强，生态越强、用户越多”的闭环。

这套规模飞轮一旦运转起来，竞争格局就彻底固定了。新进来的玩家就算有技术，也没有你的用户规模，拿不到这么低的算力价格，没有这么多数据做优化，成本永远比你高；头部大厂成本结构固化，组织庞大笨重，根本没法跟着你降价，只能慢慢让出市场。最后，你会成为通用词元市场绝对的领军者，掌握定价权，拥有稳定持续的现金流，享受规模带来的长期收益。

从长远来看，低成本词元的最终价值，是让AI词元变成像电力、网络般人人用得起的基础服务，推动整个AI行业普及。而走这条路的你，会从价格优势起步，靠规模放大优势，靠生态锁定优势，最终垄断通用词元市场。这充分说明，在词元经济时代，把规模和效率做到极致，就能做出最稳、最长久、最有壁垒的生意。规模化不是最终目的，而是通过规模建立壁垒、通过效率锁定利润，最终实现长期、稳定、可持续的商业收益。

第十四章 高价值专业词元：赚高利润长单

在通用词元市场中，价格战的硝烟已经弥漫整个赛道，从早期的每百万词元2美元的定价，到如今部分厂商已经将价格压到了0.1美元的水平，整个通用词元的市场已经进入了极致的内卷阶段。

对于选择通用词元路线的玩家来说，他们必须依靠极致的规模化来赚取微薄的利润，他们需要处理万亿级的词元调用量，才能获得百万级的净利润，这是典型的苦生意：要投入大量的资金做算力的规模化，要投入大量的人力做客户的规模化，最终却只能赚取微薄的辛苦钱，一旦市场的价格再降一点，他们就会陷入亏损的困境。

如果不想陷入这样的价格战，不想做这种规模化的苦生意，那么高价值专业词元的路线，就是最好的选择。你不需要去做大规模的用户覆盖，不需要去打惨烈的价格战，只需要聚焦一个高付费的垂直行业，做透这个行业的专业词元，就能赚取比规模化玩家还要高的利润，而且这些利润，都是长期的、稳定的，不需要持续地投入新的资源去获客，生产线一旦搭建完成，就会变成一台自动印钞的机器，持续地为你创造稳定的现金流。

高价值专业词元的核心逻辑，是价值定价，而非成本定价。对于通用词元来说，它是标准化的商品，所有厂商提供的词元没有本质的区别，所以用户只会对比价格，成本决定了最终的定价，成本越低，给出的价格就越低，最终整个行业就会陷入价格战。

这里的核心逻辑其实很简单，普通的通用词元，价格高低看成本，成本越低，卖得就越便宜，赚的都是薄利；但专业词元不一样，它的定价规则依靠价值，能给客户创造多大的价值，就能定多高的价格。客户真正在乎的，是词元能帮他们解决多大的问题、省多少钱、赚多少钱，只要能帮他们解决核心痛点，哪怕价格高一点，他们也会心甘情愿地付费。

这就是专业词元的魅力，它让你摆脱了成本的限制，摆脱了价格战的内卷，让你能够依靠价值，赚取极高的利润。

14.1 行业定向：锁定高付费的黄金赛道

做专业词元的第一步，就是选对行业，这是整个路线的核心基础，选对了行业，你就成功了一半。你要选择那些高付费、高刚需、高数据壁垒的行业，这些行业，就是专业词元的黄金赛道，它们有足够强的付费能力、有足够刚性的需求，而且有足够高的壁垒，让别人很难和你竞争。

根据全球范围内的产业实践，有4个行业是最适合做专业词元的，分别是医疗诊断、法律文书、工业质检、金融风控，这4个行业不仅付费能力极强，需求极为刚性，而且专业壁垒极高，是专业词元市场中最大的4块蛋糕。

1．医疗诊断：生命相关的极致价值

第一个赛道，就是医疗诊断，这是跟生命挂钩的生意，价值直接拉满。我们都知道，医疗行业从来都不缺花钱的地方，不管是医院还是患者，只要能保住健康、拯救生命，多少钱都愿意花。就像家里人生病住院，哪怕花光积蓄，也要找最好的医生、用最好的设备，医疗领域消费者的付费意愿，从来都是所有行业里顶尖的。

而医疗专业词元可不是随便训练出来的，它背后是数百万份三甲医院的临床病历、数千万份医学文献，还有数亿份医学影像数据，相当于把全国顶尖医生的经验和知识，都浓缩进了词元里。医生原来分析一份病历，可能要花费2小时时间，还要反复核对，生怕出错；有了医疗专业词元，5分钟就能完成，而且准确率还能提升30%以上。除此之外，它还能辅助医生做诊断、分析病情，甚至助力药物研发，这些能力，直接能帮助医院节省大量的人力成本，还能提高诊断效率，甚至能在关键时刻拯救生命，这样的价值，根本无法用金钱来衡量。

而且现在人口老龄化越来越严重，不管是老人的慢病管理，还是日常的疾病诊断，医疗需求只会越来越大，对应的医疗专业词元市场也会跟着不断扩大。你不用贪多求全，不用做所有医疗领域，只要聚焦一个细分方向，如肿瘤诊断、慢性病监测，把它做透，就相当于拥有了一个永远不会枯竭的市场。只要服务好几家医院客户，他们就会持续给你付费，而且随着他们的业务扩张、需求增加，付费金额也会越来越高，根本不用愁后续的收益。

2．法律文书：合规优先的高溢价

第二个高付费赛道，是法律文书，核心就是“合规”二字，只要能保证合

规，价格再高，企业也愿意买单。我们都知道，法律这行最忌讳的就是出错，一份合同、一份法律文书，哪怕一个字、一个条款出错，都可能给企业带来极大的损失，甚至还要承担法律责任。所以对企业来说，与其冒着风险自己起草、审核，不如花高价找专业的人、专业的工具来做，只要能保证准确、合规，钱从来都不是问题。

法律专业词元的训练基础，同样足够扎实——千万份法院的生效判决书、百万份专业法律合同，还有十万份各行各业的合规标准，把这些数据整合起来训练出的词元，比很多普通律师都专业。它能帮律师起草合同、分析案例，还能做合规审查，原来一个律师写一份复杂合同，可能要花一整天时间，还要反复修改核对，有了法律专业词元，10分钟就能生成一份合规、严谨的合同，准确率还比人工高，大大节省了律所的人力成本，也降低了出错的风险。

而且随着市场越来越规范，不管是大企业还是中小企业，对合规的需求都会越来越高，法律专业词元的市场也会持续扩大。只要深耕法律领域，熟悉某个细分方向，比如企业合规、合同起草，就能拥有稳定的客户群体。律师事务所一旦用惯了你的词元，就不会轻易更换，因为更换意味着要重新适应、重新审核，成本太高，他们会持续给你付费，而且随着律所业务量的增长，你的收益也会跟着水涨船高。

3．工业质检：生产效率的直接提升

第三个赛道，也是整个专业词元市场最大的赛道——工业质检。工业领域的场景太多了，不管是3C电子、汽车制造，还是食品加工、新能源，每个行业都有自己的质检需求，而且工业企业的付费能力都很强，只要你的词元能帮他们提升生产效率、降低成本，他们根本不会在乎价格。

工业专业词元的训练数据，更是海量——上亿张工业质检图片、千万份工业生产数据，还有百万份行业生产标准，相当于把工业生产中所有可能出现的问题、所有的质检标准，都装进了词元里。工厂的生产线原来需要几十上百名工人，每天盯着产品做质检，不仅效率低，还容易出错，次品率居高不下；有了工业专业词元，就能实现自动化质检，不管是产品的细微瑕疵，还是生产过程中的异常，都能快速识别，不仅节省了大量人力成本，还能把次品率降至最低，直接提升生产效率。

可能很多人不知道，全国有超过上千万家工业企业，每个企业都有质检、生产优化的需求，这可是一个万亿级的大市场。你不用想着做所有

工业领域，只要聚焦一个细分方向，比如3C电子产品的外观质检、汽车零部件的缺陷检测，把它做透，就能赚得盆满钵满。例如，你只服务几家大型工厂，一家大型工厂每天的词元调用量就超过1亿次，按照工业词元的定价，你每个月的收入就能超过30万元，比服务1万个个人开发者赚的钱还要多，而且客户稳定，不用反复拓客。

4．金融风控：风险控制的核心需求

第四个高付费赛道，是金融风控，这直接关系到金融机构的生死存亡，付费意愿自然拉满。金融行业最怕的就是风险，不管是银行的贷款坏账、信用卡欺诈，还是理财平台的风险管控，一旦出问题就会造成巨大的损失，所以金融机构愿意花高价寻找能帮他们降低风险的工具，只要你的词元能帮他们守住风险底线，价格再高他们也愿意接受。

金融专业词元的训练数据，更是堪称“豪华”——10亿条金融交易数据、数千万份用户信用数据，还有数百万份金融风控案例，覆盖了金融领域的所有风险场景。它能帮银行做风险评估，判断借款人的信用状况，降低坏账率；能帮金融机构做反欺诈，识别虚假交易、恶意套现；还能做智能投资顾问，给用户提供合理的理财建议。例如，一个风控词元，能把银行的坏账率从1.5%降至0.5%，对一家中型银行来说，这每年就能减少数亿元的损失，这样的价值，他们怎么可能舍不得花钱？

随着金融市场的不断发展，金融机构的风控需求只会越来越严格、越来越细致，金融专业词元的市场也会持续增长。只要做透金融领域，哪怕只服务一家中型银行，他们每天的词元调用量就超过10亿次，按照金融词元的定价，你的年收入就能超过千万，比服务10万个个人开发者还要赚钱，而且金融机构的黏性极强，一旦合作，就会长期稳定付费。

14.2 分级定价：匹配不同用户的需求

说完了4个黄金赛道，还有一个关键问题——怎么定价？

做专业词元，千万不能只定一个价格，因为用户的需求是不一样的，有的用户预算少、需求简单，有的用户预算高、需求极致，如果只定一个价格，要么会失去预算少的中小型用户，要么会浪费预算高的大型用户的付费潜力。所以最好的方式，就是做分级定价，针对不同的用户、不同的需求，定不同的价格，既能覆盖所有用户，又能最大化利润，还能让用户跟着你一起成长，持续给你付费。

具体来说，可以把专业词元分成3个等级，就像我们平时买手机，有基础版、标准版、旗舰版，不同版本对应不同的需求和价格，每个人都能找到适合自己的版本。

1．普通词元：覆盖中小型用户的入门需求

第一个等级，是普通词元，主要满足中小型用户的入门需求。如一些小型律所、个体医生、小工厂，他们的预算不多，需求也比较简单，不需要极致的准确率和合规性，只要能满足日常的基础需求就可以了。普通的法律咨询、简单的文档处理、基础的质检分析，这些需求使用普通词元就完全能够满足。

普通词元可以定价为每百万词元10美元，这个价格比通用词元高，但比专业词元低很多，对中小型用户来说，完全能承受。而且你的普通词元，比通用词元质量高很多，针对性也更强，如普通的法律词元比通用词元更懂法律术语，普通的质检词元比通用词元更懂工业场景，所以中小型用户愿意花这个钱，以获得更好的体验。

这个等级的词元，核心目的不是赚钱，而是吸引中小型用户进来，降低他们的使用门槛。这些中小型用户，现在付费虽然少，但都是你的潜在大客户，等他们的业务发展起来，需求变高了，就会升级到更高等级的词元服务，给你付更多的钱，相当于为你积累了长期的客户资源。

2．专业词元：服务中型企业的核心需求

第二个等级，是专业词元，也是你的主要收入来源，主要满足中型企业的核心需求。例如，中型律所、中型医院、中型工厂，他们的预算中等，需求也比较专业，需要足够的准确率和合规性，以满足日常的业务需求。合同起草、精准的质检分析、基础的风控评估，这些都需要专业词元服务来实现。

专业词元可以定价为每百万词元100美元，是普通词元的10倍。这个价格对中型企业来说，完全能承受，而且他们能通过专业词元，节省大量的人力成本、提高工作效率，甚至降低风险，带来的价值远比付出的价格高，所以他们愿意买单。

中型企业的数量是最多的，而且需求稳定，只要能保证词元的质量，他们就会持续付费，不会轻易更换服务商。例如，一家中型律所，每天都

需要起草合同、分析案例，只要你的专业词元能帮他们节省时间、降低风险，他们就会一直用你的词元服务，成为你长期稳定的收入来源。

3．旗舰词元：满足大型企业的极致需求

第三个等级，是旗舰词元，针对大型企业的极致需求，也是你的高利润来源。大型医院、大型律所、大型工厂、大型银行，他们的预算很高，需求也极其苛刻，需要极致的准确率和合规性来满足核心业务的需求，而且他们愿意为这份极致的价值，支付极高的费用。

旗舰词元可以定价为每百万词元1000美元，是专业词元的10倍，是普通词元的100倍。可能有人会觉得这个价格太高，但对于大型企业来说，这根本不算什么。一家大型医院，用旗舰词元辅助医生做疑难病症诊断，能提高诊断准确率，减少医疗事故，带来的价值动辄上亿，每百万词元1000美元的价格，对他们来说只是九牛一毛；一家大型银行，用旗舰词元做风控，能降低上亿元的坏账损失，花这点钱根本不值一提。

而且大型企业的黏性极强，一旦用了你的旗舰词元，就不会轻易更换，因为更换意味着要重新适配系统、重新培训员工、重新审核数据，成本太高，他们根本承受不起。所以只要你能服务好几个大型企业客户，就能获得稳定的高利润，相当于拥有了几棵“摇钱树”。

这样的分级定价，既能让中小型用户用得起，又能让中型企业愿意付更多的钱，还能让大型企业支付高溢价，覆盖了所有用户群体。更重要的是，它给用户提供了一条成长路径，中小型用户成长为中型企业，就升级到专业词元；中型企业成长为大型企业，就升级到旗舰词元，会持续支付更多的费用，让你的收益越来越稳定、越来越高。

14.3 做行业标准：锁定复购与长期合同

当然，做专业词元的终极目标，不是靠分级定价赚差价，而是制定行业标准，把你的词元变成这个行业的“规矩”，这样你就能锁定所有用户的复购，锁定长期合同，建立别人无法撼动的壁垒，让自己成为这个行业的“老大”。

可能有人会问，制定行业标准有这么重要吗？答案是肯定的，标准就是壁垒，就是垄断的“钥匙”。一旦你制定了行业标准，那么所有的企业，无论是大型企业还是小型企业，都必须用你的词元，因为不用的话，他

们的产品和服务就可能不符合行业标准，就会有合规风险，就无法正常开展业务。就像我们平时购买的电器，必须符合国家的安全标准，否则就不能上市销售，专业词元也是一样，一旦你的标准成为行业标准，别人就根本没法和你竞争，你就垄断了整个行业的市场。

那怎么制定行业标准呢？其实并不复杂，最直接的方式，就是和行业协会合作。例如，做法律专业词元，就和律师协会合作，推出法律词元的合规标准；做医疗专业词元，就和医疗协会合作，推出医疗词元的准确率标准；做工业专业词元，就和工业协会合作，推出工业词元的质检标准。把你的标准，变成公认的标准，这样所有的企业，都只能用你的词元，因为只有你的词元，才符合行业标准，才能让他们顺利开展业务。

除了制定行业标准，还有一个关键动作，就是做深度的系统集成。把你的词元和行业的ERP系统、业务系统打通，深度融入企业的日常运营中。例如，一家工厂把你的工业词元集成到他们的生产系统里，所有的质检、生产优化，都通过你的词元来完成；一家律所，把你的法律词元集成到他们的办公系统里，所有的合同起草、案例分析，都离不开你的词元。这样一来，企业一旦用了你的词元，就没法更换了，因为更换意味着要重新改造整个系统，花费大量的时间和金钱，成本太高。

通过这两种方式，你就锁定了用户的迁移成本，用户一旦用了你的词元，就再也离不开你了。这时候，你就可以和企业签订长期合同，比如3年、5年，甚至10年，锁定未来几年的稳定收入，你的现金流会非常稳定，公司的估值也会大幅提升，因为你有确定的、长期的收益。

到了这一步，你就拥有了整个行业的用户，不需要再花时间寻找新客户，不需要再打价格战，不需要再做规模化的苦生意，只要服务好现有的用户，他们就会持续给你付费。你的词元生产线，就变成了一台自动“印钞机”，只要开着，就会有源源不断的钱流进你的口袋，这就是高价值专业词元的终极形态，也是这个AI时代最轻松、最赚钱的生意。

最后要提醒一句，这条路线最适合那些有行业资源的人。如果你原来就是医疗、法律、工业或者金融行业的从业者，手里有行业资源、有行业数据，那你入局的优势会非常大，不用从零开始积累，只要把这些资源利用起来，搭建专业词元生产线，就能快速实现盈利，抓住词元经济的高利润红利。

第十五章 词元出海：中国算力的数字化出口

词元出海，是中国词元玩家最大的增量机遇。这是因为中国在词元生产领域已经拥有了全球领先的优势，我们的低成本、高效率在全球市场是碾压级的，我们可以把中国的词元卖给全球的用户，把中国的算力变成数字化的出口商品，赚取全球化的收益。

过去中国的出口商品主要是衣服、鞋子、电子产品，是实物的出口，而未来中国的出口商品是词元，是数字的出口，是AI时代的新的大宗商品。这个市场，比实物出口的市场还要大，还要有潜力。

15.1 全球需求爆发：中国词元的碾压级优势

全球的词元需求正在爆发式地增长。根据36氪的报告，2026年中国模型在OpenRouter的调用量，已经占到了50%，把OpenAI、Anthropic这些巨头都甩在了后面，中国的词元已经成为全球市场的主流。MiniMax、月之暗面、智谱、阶跃星辰、DeepSeek、小米等中国公司的基础模型Token调用量轮番冲进全球前三。

这说明全球的用户都需要更便宜、更高质量的词元，而中国的词元在全球市场拥有碾压级的优势。

1．成本优势

中国有全球最便宜的绿电、最便宜的算力和最高效的模型，我们的词元生产成本只有海外的1/10～1/6，但是服务能力却差不多。这样的性价比，全球的用户根本无法拒绝。

2．模型优势

中国的开源模型发展得非常快，Qwen、MiniMax、DeepSeek这些模型能力已经接近GPT-4，但是它们是开源的，我们可以做微调、做优化，成本更低、效率更高。

3．工程师优势

中国有全球最多的AI工程师，我们的优化能力、推理效率比海外玩家高

很多，我们能够把硬件的潜力挖掘到极致，把词元的成本降至最低。

中国词元出海的极致性价比碾压了美国的模型，现在，在东南亚、日韩、拉美这些词元使用量增长最快的地区，中国的词元已经占据了70%的市场份额，未来我们还会进入欧美市场，抢占更多的市场份额。

15.2 交付通道：把中国词元送到全球用户手里

很多人会问，如何将中国的词元送到全球用户的手里？网络延迟怎么办？跨境的问题怎么办？

实际上，现在已经有成熟的交付通道，既不需要自己建海外机房，又不需要自己做跨境网络，只要用现成的CDN、跨境专网、海底光缆，就能把你的词元低延迟地送到全球用户的手里。

首先，可以用Cloudflare的CDN，把API的请求加速分发到全球用户。这样，哪怕你的服务器在国内，海外用户调用你的API延迟也只有几十毫秒，与本地服务的延迟差不多。

其次，很多云厂商都有跨境专网，如阿里云的全球专网、腾讯云的跨境专线，能够把你的数据通过专用线路传到海外，延迟低，稳定性高，不会有网络波动的问题。

现在海底光缆已经非常发达了，中国到美国的海底光缆延迟只有100多毫秒，完全能够满足词元推理的需求，用户根本感觉不到延迟。

另外，还可以用OpenRouter这样的聚合平台，可以帮你做好所有的交付通道。它们有全球的节点和网络，只要把API接到它们的平台，它们就会把你的词元分发到全球的用户，帮你加速和保障稳定性，而你根本不需要考虑任何交付的问题。

这些平台解决了中国模型走向全球的“最后一千米”的问题。它们整合了多家中国模型厂商的API，作为一个统一的入口向全球（特别是北美、亚太等地区）的客户提供服务。这种模式的优势在于，能够帮助中小型模型厂商快速触达海外市场，同时为海外用户提供更加多元、更具性价比的选择。

15.3 结算与合规：全球化的变现通道

很多人还会问，海外的用户怎么付费？外汇的问题怎么解决？合规的问题怎么解决？

这些问题早就解决了。OpenRouter这样的平台会帮你做好结算，帮你收取海外用户的美元，然后兑换成人民币汇入你的账户。

合规的问题也不用担心。词元出海是数字服务的出口，是中国数字产品的出口，国家是支持的，而且只要通过正规的平台结算，都是合法合规的，不会有任何的风险。

现在已经有超过400个模型通过OpenRouter这样的平台提供词元调用服务，它们的收入都是合法合规的，都是正常的数字服务，没有任何问题。

词元出海的收入是以美元计价的，你还能享受汇率的红利。比如，在人民币贬值的时候，用美元兑换人民币会让到手的人民币变多，这也是额外的收益。

现在，词元出海已经成为中国数字出口的新的增长点。根据海关数据显示，2025年中国数字服务的出口量增长了30%，其中与词元相关的出口量增长了500%，成为最大的增长点。

未来词元出海的市场会越来越大。根据预测，到2030年，全球词元市场的规模会达到3万亿美元，中国的词元出口会占到30%，也就是9000亿美元，这比我们现在所有数字出口的总和还要大。

第十六章 智能体时代的词元爆发：**OpenClaw**带来的增量

未来智能体（Agent）时代的词元市场，会比现在大10倍甚至100倍，OpenClaw引爆的智能体革命，会带来词元需求的指数级爆发，这是词元商业化的终极增量，也是从事词元服务的企业最大的机遇。

16.1 OpenClaw引爆Agent：词元消耗的10倍级增长

2026年，OpenClaw的发布引爆了智能体的革命。智能体，就是能够自主完成任务的AI，它不需要一步一步的指令，能够自己理解目标，自己做计划，自己调用工具，自己完成任务。比如，让智能体做一个市场调研，它会自己去搜索网络、整理数据、做分析、写报告，整个过程它自己就完成了。

但是，这个过程会消耗海量的词元。因为智能体每做一步，都要调用模型，都要消耗词元。一个智能体任务消耗的词元，是传统人机对话的100～1000倍。

黄仁勋在演讲里提到，ClaudeCode的出现彻底改变了词元的消耗模式，原来一个工程师一天消耗1万词元，现在用了ClaudeCode，一天会消耗100万词元，增长了100倍。

根据研究报告数据显示，OpenClaw发布之后全球的词元调用量在3个月里增长了10倍。这是因为智能体的需求爆发了。原来一个用户一天调用10次模型，现在一个智能体一天调用1000次模型，词元的消耗量直接增长了100倍。

比如，一个写小程序的简单智能体任务，需要读取用户需求、写代码、编译、测试、修改，整个过程要调用几十次模型，消耗几十万词元，这比与ChatGPT聊10分钟消耗的词元多了几十倍。

更复杂的任务，比如做一个年度的财务分析，智能体要读取所有的财务报表、要做分析、要做图表、要写报告，整个过程要调用几百次模型，消耗几百万词元，这是传统模式根本没法相比的。

这就是智能体带来的增量，它把词元的需求从线性增长变成了指数增

长，原来生产线一天生产10亿词元就够了，现在一个智能体客户一天就要消耗10亿词元。

16.2 工程师词元预算：企业的刚性支出

随着智能体的普及，企业的词元预算正在变成刚性支出。就像现在的水电费、网费一样，每个企业、每个工程师，都要有固定的词元预算，这是最稳定的需求。

现在很多企业都需要给工程师配词元预算，每个工程师每个月100万词元的预算是标配。这是因为工程师要用智能体写代码、做测试、做分析，没有词元，他们就无法工作。

比如，Google现在给每个工程师每个月配200万词元的预算、Microsoft给每个工程师配150万词元的预算、Meta给每个工程师配300万词元的预算。这些预算都是刚性的，不管价格怎样他们都要支付。没有词元，工程师的工作效率就会下降，企业就失去了竞争力。

根据统计，全球有3000万开发者，如果每个工程师每个月消耗100万词元，那就是每年360万亿词元的需求，这比现在全球的年调用量多10倍。

这个需求是刚性的，是稳定的，不会因为经济周期波动，不会因为市场变化就消失了。智能体已经变成了工程师的必备工具，就像电脑和互联网，没有它就无法工作。

现在很多硅谷的创业公司在招聘员工时都会承诺给员工配置词元的预算，这已经变成了福利，变成了吸引人才的手段。

对于词元生产者来说，这意味着有了稳定的、海量的、刚性的需求，只要把词元卖给这些企业，卖给这些工程师，就有了永远接不完的订单，生产线永远不会缺用户，生产者也永远不会缺收入。

16.3 抓住智能体生态：抢占下一代词元消费入口

智能体时代最大的机遇，就是抢占下一代的词元消费入口。原来的词元消费入口是OpenAI的API，是Anthropic的API；而未来，词元消费入口是智能体平台，是OpenClaw，是这些智能体的基础设施。

谁能抢占这个入口，谁就能掌控未来的词元市场。所有的智能体都会通过这个入口消耗词元，所有的需求都会通过这个入口，流向词元生产者。

比如，OpenClaw现在已经成为全球最大的智能体平台，所有的智能体开发者都在OpenClaw上开发智能体，所有的企业都在OpenClaw上使用智能体，他们的词元消耗都要通过OpenClaw。OpenClaw就成为词元消费的入口，它们掌握了所有的需求。

如果你能和OpenClaw合作，把你的词元接入OpenClaw的平台，那么所有的智能体用户都会变成你的用户，所有智能体的词元消耗都会变成你的收入。你根本不需要自己找用户，只要接入这个入口，就会有海量的需求。

比如，Fireworks AI就是最早接入OpenClaw的词元厂商，现在他们的大部分收入都来自OpenClaw的智能体用户。

MiniMax也接入了OpenClaw，现在他们的智能体相关收入已经占到了总收入的50%，而且还在快速增长，未来会占到80%以上。

智能体生态会成为词元经济的核心，所有的词元需求都会来自智能体，所有的词元生产都会围绕智能体。谁能抢占这个生态的入口，谁就能掌控未来的词元市场，谁就能拥有最高效的“印钞机”。

根据预测，到2030年，智能体带来的词元需求会占到整个词元市场的80%，也就是说，未来的词元经济几乎都是智能体带动的，现在抓住智能体的机遇，就是抓住了词元经济的未来。

到这里，我们已经把词元商业化的所有路径都讲完了，从上游的五大商业模式，到低成本的规模化增长，到高价值的垂直利润，到词元出海的全球化机遇，再到智能体时代的增量爆发，你已经知道了怎么把你的生产线，变成一台自动运转的“印钞机”。

词元经济的商业化是AI时代最容易、最可行的创业，也是最赚钱的创业，不需要大量的资金、庞大的团队和复杂的推广，只要搭建好生产线，选对商业模式，就能赚到持续的、复利的、被动的收入。词元经济的浪潮才刚刚开始，智能体的爆发、词元出海的机遇、垂直行业的需求，都在打开巨大的市场，现在就是最好的进入时机。

第五部分

未来趋势：
词元经济的终极格局

当我们了解了词元生产和商业化的全流程，不禁要问，词元经济的未来会是什么样子？它的终极格局，会如何改变整个世界的经济、产业、财富的逻辑？

在这一部分中，我们将站在未来的视角预判词元经济的长期趋势，从词元成为基础能源的必然，到全球产业链的格局重构，再到下一代词元工厂的形态，以及个人与企业的终极财富形态。最后梳理在这股浪潮中投资与布局的优先级，帮助创业者抓住这千载难逢的机遇。

第十七章 词元成为AI时代基础能源：像电力、石油一样成为刚需

黄仁勋曾经说过一句振聋发聩的话：“词元，就是AI世界的度电。”这句话，揭示了词元经济的终极未来：未来，词元会像电力、石油一样成为整个社会的基础能源，成为所有智能活动的结算单位，成为整个经济体系的底层基础设施。

这个趋势，已经在快速变为现实。现在我们已经能看到，词元正在渗透到社会的每一个角落，从个人的日常使用，到企业的生产经营，到工业的制造流程，到政府的政务服务，所有的智能活动都离不开词元，所有的价值结算都离不开词元，词元正在成为AI时代新的基础能源。

17.1 从工具到基础设施：词元的全民渗透

10年前，互联网还是一个工具，是一种可选的生活方式。而现在，互联网已经成为基础设施，没有它就无法工作，无法生活，无法与世界连接。

词元正在走同样的路。5年前，AI还是一项实验室技术，词元还是一个计算机专业名词，而现在，词元已经开始渗透到普通人的生活里，未来它会像互联网一样成为全民的基础设施。

首先，是个人生活的渗透。现在普通人用AI聊天、用AI写文案、用AI做翻译、用AI做规划，这些都要消耗词元。未来家里的智能设备，如智能音箱、智能电视、智能汽车、智能机器人，都要用到AI，都要消耗词元，词元会成为家里新的“水电煤”。以后每个月除了要交电费、交网费，可能还要交词元费。

比如，现在已经有很多家庭每个月的词元消费超过了100元，未来这个数字会涨到500元，甚至1000元，因为要使用的智能服务越来越多，需要的词元也会越来越多。

其次，是企业经营的渗透。现在企业用AI做客服、做营销、做研发、做风控，这些都要消耗词元，未来企业的所有业务流程、管理流程、决策流程都要用到AI，都要消耗词元。词元会成为企业的核心运营成本，就

像现在的场地费和人力成本一样。

根据预测，到2030年企业的词元成本会占到企业营收的5%。也就是说，一个营收10亿元的企业每年要花5000万元购买词元，这是刚性的支出，没有词元企业的效率就会下降，企业就失去了竞争力。

最后，是工业生产的渗透。现在工业AI已经开始落地，质检、生产、调度、运维都要用到词元。未来整个工业的生产流程都会被AI重构，所有的生产环节都要消耗词元，词元会成为工业的新动力，就像当年的电力、石油一样推动工业的发展。

17.2 每瓦词元数：国家与企业的核心**KPI**

随着词元成为基础能源，一个新的核心指标正在成为国家、企业、数据中心的核心KPI，这个指标就是黄仁勋提出的每瓦词元数（Tokens per Watt）。

在固定的功率限制下，谁的每瓦词元数吞吐量最高，谁就拥有最低的生产成本，谁就能在词元经济中获胜。过去，衡量一个数据中心效率的是PUE，也就是电源使用效率，衡量用了多少电，有多少电用于服务器，多少电被浪费。而未来，衡量一个数据中心的效率是每瓦词元数，衡量一度电能生产多少词元。这个指标，才是核心。

对于国家来说，每瓦词元数是国家竞争力的核心，谁的每瓦词元数越高，谁的词元成本就越低，谁就能在全球的词元市场中拥有竞争力，谁就能出口更多的词元，赚更多的钱。

中国现在的每瓦词元数，已经是全球最高的。因为我们有最便宜的绿电、最高效的模型、最强大的优化能力。中国的每瓦词元数，能达到美国的2倍、欧洲的3倍，这就是我们的核心竞争力，这也是为什么我们的词元能够在全球市场碾压对手。

对于企业来说，每瓦词元数是企业盈利能力的核心，每瓦词元数越高，词元成本越低，利润就越高，就能在价格战中获胜，就能占领市场。

未来，每瓦词元数会像GDP一样，成为衡量一个国家、一个企业的核心指标。所有的技术创新和产业升级，都会围绕提升每瓦词元数来展开，这就是词元经济的核心。

17.3 标准化的价值度量衡：统一的结算体系

词元成为基础能源的另一个标志，就是标准化的价值度量衡，即统一的结算体系。未来所有的AI服务、智能活动，都会用词元来计量、来结算。就像现在的电力，一度电就是一度电，不管是用来照明，还是用来做饭，还是用来进行工业生产，它的计量都是统一的，结算也是统一的。

过去AI服务的计费是混乱的。有的按调用次数计费，有的按时长计费，有的按功能计费，没有统一的标准，用户不知道自己花钱买到了什么。而现在，词元成为统一的计量单位，不管是什么模型、什么服务，都按词元计费，用户可以清晰地知道自己的成本是多少，自己的消费又是多少。

这个标准化还会进一步深化，词元的计量标准会成为国际标准。无论是中国的模型，还是美国的模型，无论是通用的模型，还是专业的模型，词元的计量都是统一的，结算也是统一的。这样，词元才可以像电力一样，自由地流动、交易和分配。

比如，现在用OpenAI的词元和用MiniMax的词元，计量是一样的，可以自由地切换和选择，就像可以选择不同电力公司的电一样，这就是标准化的力量。

词元的标准化，还会催生新的金融产品。比如，词元期货、词元期权、词元交易市场，就像石油、电力一样，你可以提前锁定词元的价格，可以交易词元的产能，可以做词元的金融衍生品，这会让词元经济变得更加成熟、更加完善。

第十八章 全球词元产业链格局：中国成为最大供给国

随着词元经济的全球化，全球的产业链格局正在被重构。过去AI产业链是由美国主导的，使用的是美国的模型、美国的算力、美国的服务。而现在，中国正在崛起，开始成为全球最大的词元供给国，甚至中国还会成为全球的词元产业中心。

18.1 产业链全景：生产、优化、流通、应用

词元经济的产业链分为4个核心环节，分别是生产、优化、流通、应用，这4个环节构成了完整的价值链条。中国在这4个环节都已经拥有了全球领先的优势。

首先是生产环节，也就是词元的生产。这个环节是产业链的上游，核心是算力、电力、模型，中国在这个环节，拥有全球领先的优势，有着全球最大的算力集群、最便宜的绿色电力和最高效的开源模型。中国的词元生产成本只有美国的1/10～1/6，这是核心竞争力。

其次是优化环节，也就是词元的微调、定制。这个环节是产业链的中游，核心是技术、数据、工程师。中国有全球最多的AI工程师，优化能力和推理效率是全球最高的，能够把硬件的潜力发挥到极致，把词元的成本降至最低，垂直模型和微调技术也是全球领先的。

再次是流通环节，也就是词元的分发、交付、结算。这个环节是产业链的中游，核心是网络、平台、渠道。中国已经建立了完善的流通体系，现有的CDN、跨境专网和模型市场能够把中国的词元高效地送到全球的用户手里。结算体系也已经成熟，能够帮助中国的词元厂商收取全球的费用。

最后是应用环节，也就是词元的使用。这个环节是产业链的下游，核心是场景、用户、需求。中国有全球最大的AI应用市场，开发者、用户和场景数量也最多。中国的应用创新是全球最快的，词元需求也是全球最大的。

这4个环节中国都已经领先了，所以全球的词元产业链正在向中国转移。过去全球的AI产业链是美国主导的，现在形成了中美双主导的格

局，中国正在成为全球的词元工厂。

18.2 涨价传导链：全球需求下的全链重估

随着全球词元需求的爆发，全球算力开始短缺，全球电力开始紧张。这就会带来涨价的传导链，从下游的词元，到中游的算力，到上游的电力，整个产业链都会被重估，而中国作为全球最大的词元供给国，将会成为最大的受益者。

假设到2027年，全球的算力市场规模达到1万亿美元，而供给根本跟不上。于是算力会短缺，价格会上涨，然后电力的需求上涨，电力的价格也会上涨，整个产业链都会涨价。

这条涨价的传导链如下所述。

首先，下游的词元需求量爆发式增长。全球的用户都需要词元，需求远远大于供给，所以词元的价格开始上涨。但是中国的词元因为成本低，所以涨价的空间更大、利润更高。

其次，词元需求的增加，带动算力需求的增加。全球的算力开始短缺，算力的价格开始上涨。但是中国的算力因为成本低，所以涨价给中国算力厂商带来了更高的利润。

最后，算力需求的增加，带动电力需求的增加，全球的电力开始紧张，电力的价格开始上涨。但是中国的绿电因为成本低，所以电力的涨价给中国电力厂商带来了更高的利润。

也就是说，在整个涨价的过程中，中国企业都是最大的受益者。因为成本低、产能大，我们能够满足全球的需求，能够在涨价的过程中赚更多的钱。

这就是中国的优势。在全球涨价周期里，我们因成本低，所以有定价权，能够把涨价的压力转嫁给海外的用户，能够赚更多的钱，而海外的厂商只能自己承担成本的上涨，利润越来越薄。

18.3 全球格局的重构：从美国单极到中美竞合

随着中国的崛起，全球词元产业链的格局正在重构。过去是美国单极主

导，美国的模型、算力和服务垄断了全球市场。以后将会变成中美竞合的格局，中国主导中低端的词元市场，主导新兴市场，美国主导高端的词元市场，主导欧美市场。两者既有竞争，又有合作。

在中低端的通用词元市场，中国的厂商已经垄断了全球的市场，因为我们的性价比太高，美国的厂商根本没法和我们竞争。现在在东南亚、日韩、拉美、非洲这些新兴市场，中国的词元已经占据了大部分的市场份额，未来这个份额还会继续增长。

在高端的专业词元市场，美国的厂商还有一定优势。比如，顶尖的GPT-5，顶尖的医疗、金融专业模型。但是，中国的厂商正在快速追赶，相信用不了多久我们就会追上他们，占领高端市场。

但中美两国也会有合作，中国厂商给美国企业提供低成本词元，美国厂商给中国企业提供高端技术，两者互相补充，共同服务全球用户。

比如，OpenAI现在也在采购中国的词元，用来做他们的低端服务。因为中国的词元太便宜了，比他们自己生产的便宜很多，所以他们也愿意购买中国的词元，用来降低自己的成本。

而中国的企业也会采购美国的高端词元，用来做自己的高端服务、用来做模型的蒸馏、用来做训练。两者相互需要，相互合作。

随着中国官方正式定名“词元”，并赋予其“价值锚点”的战略地位，一场围绕词元标准制定的国际竞争也将不可避免地展开。中国将凭借其庞大的市场规模和应用场景，积极推动词元计量、定价、结算等标准体系的建立，并试图将其推广为国际标准。而以美国为首的西方国家，也可能基于其现有的技术优势，提出自己的标准方案。

未来，全球可能会形成以中美两国为核心的两大词元标准体系，即中国是全球最大的词元供给国，美国是全球最大的高端技术输出国。两者共同主导全球的词元经济，推动整个产业的发展。两国之间既有竞争，又可能在一定范围内实现互操作和互联互通。这场标准之争，将深刻影响未来全球数字经济的格局和各国的产业竞争力。

第十九章 词元工厂的下一代形态

随着词元需求量的爆发式增长，词元工厂的形态也在快速进化，从传统的IDC机房，到AI原生的词元工厂，再到下一代的超级工厂。技术的进步正在把词元工厂的效率提升到极致，把成本降至最低。

未来的词元工厂会是1吉瓦（GW）级的超级AI工厂，整个工厂就是为了生产词元而设计的。所有的技术和架构，都是围绕把每度电转化为最多的词元，这就是下一代的词元工厂的形态。

19.1 硬件的进化：液冷、**CPO**、高速互联

下一代词元工厂的硬件会有巨大的进化。传统的风冷服务器、电互联已经无法满足现实需求了。未来，液冷、CPO、NVLink这些新技术会成为标配，把硬件的效率提升到极致。

首先，是液冷技术。传统的风冷PUE很高，通常在1.4以上，也就是说，给服务器供1度电，还要用0.4度电来散热，浪费了很多的电力，而液冷技术PUE可以做到1.1以下，甚至1.05，也就是说，几乎所有的电力都用来做计算，基本没有散热浪费了。这样，电力成本会下降25%甚至更多。

同时，液冷技术能够让芯片的温度更稳定，能够让芯片超频运行，进一步提升算力，提升词元的生产效率。比如，英伟达的L4卡，使用液冷技术能够超频20%，词元的生产速度提升20%，成本降低20%。

然后是CPO，也就是共封装光学。传统的电互联带宽不够，延迟很高，功耗很高，满足不了大模型的高速互联需求。而CPO把光引擎和芯片封装在一起，带宽提升了10倍，延迟下降了90%，功耗降低了80%。这样，服务器之间的互联就会更快、更便宜，调度系统就能够更好地调度算力，更好地利用资源。

还有NVLink高速互联。英伟达新推出的NVLink-72能够把GPU之间的互联带宽提升至72GB/s，比原来的PCIe快了10倍。这样可以把上千个GPU连在一起，组成一个超级集群，用来做超大模型的推理，做大规模的词元生产，效率提升了很多。

这些硬件的进化能够把词元工厂的效率提升1倍以上，把成本降低一半以上，这就是下一代词元工厂的硬件基础。

19.2 吉瓦级超级**AI**工厂：电力直接转化为收入

有了这些硬件的基础，下一代的词元工厂会是1吉瓦级的超级AI工厂。也就是说，整个工厂的功率是100万千瓦，相当于一个小型城市的用电量。这样的工厂每天能够生产超过1000亿词元，能够服务超过1000万的用户，一年的收入能够超过1000亿美元。

黄仁勋认为，未来全球会有超过100个这样的1吉瓦级的超级AI工厂，这些工厂会成为词元生产的核心，它们会把电力直接转化为收入。用一度电，生产出相应数量的词元，就能卖出相应的钱，就像发电厂，把煤转化为电，再转化为收入，而词元工厂把电转化为词元，再转化为收入。

未来，这样的超级工厂会越来越多，预测到2030年，中国会有超过50个这样的1吉瓦级的超级AI工厂，美国会有30个，欧洲会有20个，这些工厂会生产全球90%的词元，成为词元经济的核心基础设施。

19.3 边缘词元工厂：把生产放到距用户最近的地方

除了中心的超级AI工厂，还会有边缘的词元工厂。也就是说，把小型的词元生产线部署到边缘，部署到用户的身边，部署到工厂里，部署到城市里，满足用户的本地化需求。

很多用户对数据隐私的要求很高，他们的核心数据不能传到中心的工厂里。所以他们需要边缘的词元工厂，在本地完成推理，完成词元的生产，数据不需要出本地，既满足了AI的需求，又保证了数据的安全。

还会出现城市级的边缘词元工厂。运营商会在城市的机房部署边缘词元生产线，给城市里的企业和用户提供低延迟的词元服务。比如，自动驾驶的车需要低延迟的词元服务，中心工厂延迟太高，边缘的工厂延迟只有几毫秒，就能更好地满足需求。

未来，边缘词元工厂和中心超级AI工厂会形成互补：中心工厂生产通用词元，服务全球的用户；边缘工厂生产专属词元，服务本地用户。两者共同构成词元生产的网络，覆盖所有的需求。

19.4 前沿探索：太空数据中心、主权**AI**节点、多模态

除了这些，还有很多前沿的探索，如太空数据中心。也就是说，把词元工厂放到太空里。太空里有无限的太阳能，有极低的温度，散热也不需要成本，更没有地理的限制，能够给全球用户提供最低成本的词元。

SpaceX计划在2030年部署第一个太空AI数据中心，用卫星的太阳能给GPU供电生产词元，然后通过星链把词元传给全球用户。这样的话，词元的生产成本会比地面的生产成本低90%，因为太空的电力是免费的，散热也是免费的。

还有主权AI节点。每个国家都会有自己的主权词元工厂，因为很多国家对数据主权有要求，他们的词元不能放到别的国家，所以他们会建立自己的主权AI节点、自己的词元工厂，以满足自己国家的需求。

比如，东南亚、中东和拉美地区的一些国家，他们都在与中国的企业合作，建立自己的主权AI节点。中国的企业给他们提供模型、提供硬件、提供技术，帮他们建自己的词元工厂。这样，他们就有了自己的主权AI节点，满足了他们的国家需求，而中国的企业也赚到了钱。

当前词元的概念主要与文本处理相关。但未来随着多模态大模型（Multimodal LLMs）的成熟，词元的内涵将得到极大扩展。一个词元将不再仅仅代表文本片段，也可能代表一张图片、一段音频、一段视频，甚至是一个三维空间的点云。词元经济将从一个主要处理文本的“文本经济”，演变为一个能够理解和生成跨多种媒介信息的“多模态经济”。这将催生出更丰富的应用场景，如AI视频剪辑师、AI音乐创作人、AR/VR中的智能虚拟角色等。

这些前沿的探索，会进一步推动词元工厂的进化，会把词元的成本降得更低，把词元的覆盖面做得更广，让词元能够服务全球的每一个用户。

第二十章 个人与企业的终极财富形态

词元经济的终极未来会彻底改变个人与企业的财富形态。过去，财富是房子、是股票、是现金，而未来，财富是词元产能、是词元生产线。这是因为，词元生产线能够带来永久的现金流，带来复利的增长，带来被动的收入，这是企业的终极财富形态。

20.1 个人词元生产线：人人可搭建的小型**“**印钞机**”**

过去，搭建词元生产线是巨头的专利，需要几十亿的资金，需要几千人的团队，普通人根本无法做到。未来，随着技术的进步、开源生态的成熟、模型市场的完善，个人也能够搭建自己的词元生产线，每个人都能够拥有自己的小型“印钞机”。

由于搭建词元生产线的门槛正变得越来越低，这样的案例会越来越多。现在尚需要懂一些技术，未来可能会出现零代码的搭建工具，只要点几下鼠标就能够搭建词元生产线。不需要懂任何技术，只要有自己擅长的行业数据就能做。

比如，一名教师有自己的教学数据，就可以搭建一个教育词元生产线，出售给其他教师和学生；一位医生有自己的诊断数据，就可以搭建一个医疗词元生产线，出售给其他医生和医院；一名设计师有自己的设计数据，就可以搭建一个设计词元生产线，出售给其他设计师和服装企业。

未来每个人都可以做自己的垂直词元，都可以搭建自己的小型生产线，都能够拥有自己的“印钞机”。只要把自己的知识转化为词元，就能够持续地得到收入，真正地实现知识变现，实现财富自由。

预测到2030年，全球会有超过1000万的个人词元生产者，他们会生产全球20%的词元，会成为词元经济的重要组成部分，会彻底地改变个人的财富形态。

20.2 企业词元战略：每家**SaaS**都成为**Agent**即服务公司

对于企业来说，未来所有的企业都要有自己的词元战略，所有的SaaS公司都将变成Agent即服务的公司，都会拥有自己的词元生产线，只有这

样企业才能在智能体时代拥有竞争力。

过去，SaaS公司出售软件，出售的是功能和工具。而未来，SaaS公司会变成Agent即服务的公司，会拥有自己的词元生产线，出售的是智能体，是词元服务。

Salesforce是全球最大的SaaS公司，现在已经推出了自己的Agentforce，即智能体服务。他们搭建了自己的词元生产线，给客户提供销售和客服词元，提供智能体的服务。截至2025年，这部分业务的年度经常性收入（ARR）已经接近14亿美元，占到了企业总收入的13%，未来将会更多。

国内的用友网络科技股份有限公司也推出了自己的智能体服务，搭建了自己的词元生产线，为他们的企业客户提供ERP词元、财务词元，提供智能体服务，他们的收入也在快速增长。

为什么要这么做？因为，未来智能体会成为企业的核心工具，所有的用户都会用智能体来完成工作，而智能体需要消耗词元。如果你是SaaS公司，拥有自己的词元生产线，你才能够做到更低的成本，获得更高的利润，你才能够给用户提供更好的服务，锁定他们，建立你的壁垒。

未来所有的SaaS公司都会变成Agent即服务的公司，都会拥有自己的词元生产线，这是企业的终极形态。只有这样，他们才能够在智能体时代生存下来、发展起来。

20.3 持有词元产能=持有永久现金流

不管是个人还是企业，终极的财富形态就是持有词元产能。词元产能能够带来永久的现金流，只要生产线在，只要词元在，就能够持续地赚钱，拥有永久的现金流。

过去，你买房子、收房租，这是现金流。但是，房子会折旧、会老化、会有政策的风险，而且房租的增长很慢。

过去，你买股票、收分红，这是现金流。但是，股票会波动、会有退市的风险，而且分红的增长不确定。

但是，词元产能不一样，只要维护好词元生产线，就能够永久地产生现

金流，而且随着词元需求量的爆发式增长，随着词元价格的上涨，收入会呈指数级增长，现金流会越来越多、越来越稳定。

未来，词元产能会成为最优质的资产，比房子、股票及其他任何资产都要优质。因为它的现金流最稳定、增长最快、风险最低，所以持有词元产能，就是持有了终极的财富。

未来词元产能会成为企业和个人的核心资产，谁拥有的产能越多，谁的财富就越多，谁的竞争力就越强，这就是词元经济的终极真相。

第二十一章 投资与布局优先级

面对词元经济的巨大浪潮，很多人都会问，我该怎么投资？我该怎么布局？哪些机会是最确定的？哪些机会是最大的？

根据产业的发展规律，词元经济的投资与布局是有优先级的，不同的阶段会有不同的机会。我们按照时间的顺序，把机会分为短期、中期、长期和终局4个阶段，梳理最确定的布局方向。

21.1 短期：存储与显存，弹性最大的机会

短期，也就是未来1-2年，最确定的机会是存储与显存。因为词元需求的增长首先带来的是存储和显存的短缺，这部分的弹性是最大的。

因为大模型的推理需要大量的显存，来存储模型的权重和KV缓存，词元的需求越大，需要的显存就越多。而且，推理的过程还需要大量的存储来存储数据、存储用户的请求、存储中间的结果。所以，词元需求的增长首先带动的就是存储与显存的需求。

过去两年推理的需求增长了100倍，而显存的供给只增长了10倍，所以显存的短缺非常严重，显存的价格在过去的一年里涨了3倍，而且还在涨。

存储的需求也是一样。词元的生产需要大量的数据，需要大量的存储，过去一年存储的价格涨了50%，而且还在涨。因为，需求远远大于供给。

所以短期最确定的机会，就是存储和显存。相关的公司，比如，显存的芯片公司、存储的芯片公司、服务器的内存公司，这些公司会最先受益。他们的业绩会爆发式地增长，他们的股价也会爆发式地增长。

比如，美光的HBM显存，在过去一年里价格涨了3倍、企业营收涨了200%、股价涨了150%，这就是短期的机会。

还有，长江存储的AI存储芯片，在过去一年里价格涨了50%、企业营收涨了100%、股价涨了80%，这也是短期的机会。

短期的机会是最确定的，也是最快的。因为需求已经大量增长了，但供给还没跟上，所以相关公司的业绩会立刻兑现，弹性非常大。

21.2 中期：算力芯片与服务器，业绩确定的主线

中期，也就是未来2-5年，最确定的机会是算力芯片与服务器。因为当满足了存储和显存的需求之后，接下来就是算力需求的增长，算力芯片和服务器会成为整个产业的主线，业绩非常确定。

黄仁勋预测，到2027年全球的算力需求会达到1万亿美元，而现在全球的算力供给只有2000亿美元。照这样计算，未来5年算力的需求会涨5倍，算力缺口非常大。

所以，中期算力芯片和服务器会是最大的机会，相关的芯片公司，如英伟达、AMD、英特尔，还有国内的华为、寒武纪、海光，还有服务器厂商，如浪潮、戴尔、惠普这些公司，会受益于算力需求的爆发，他们的业绩会持续地增长。

英伟达在过去两年里企业营收增长了300%，股价上涨了200%，未来还会继续增长，因为算力的需求还在增长，他们的订单已经排到了2028年。

还有像浪潮这样的国内服务器龙头企业，在过去两年里的企业营收增长了200%，股价上涨了150%，未来还会继续增长，因为国内的算力需求也在增长，他们的订单也排到了2028年。

中期的机会是整个词元经济的主线，持续的时间会很长，业绩非常确定，未来5年都会是这些公司的黄金期。

21.3 长期：绿电、电力设备、**AIDC**，壁垒最深的资产

长期，也就是未来5-10年，最确定的机会是绿电、电力设备、AIDC。因为当满足了算力的需求之后，接下来就是电力需求的爆发，电力会成为词元生产的第一瓶颈，而绿电、电力设备、AIDC这些资产，壁垒最深、最长期、最稳定。

未来1吉瓦级的超级AI工厂会成为主流，每个工厂都需要1吉瓦的电力。全球100个这样的工厂就需要1亿千瓦的电力，这相当于整个中国的电力

装机容量的10%。所以电力的需求会爆发式地增长，电力的瓶颈会越来越明显。

绿电的机会是最大的，因为词元生产需要大量的低成本绿电，绿电的需求会爆发式增长。相关的公司，如风电、光伏、水电公司，以及电力设备公司，会受益于电力需求的爆发，业绩会长期增长，非常稳定。

还有AIDC，也就是AI数据中心，是词元生产的基础设施，壁垒非常深。因为如果要建一个AIDC，需要拿到电力指标、土地指标、能耗指标，这些都是稀缺的资源，别人根本无法复制。这些资产壁垒非常深，长期来看会越来越值钱。

例如，光环新网在呼和浩特的AIDC，他们拥有的绿电指标是稀缺的，别人根本拿不到。所以，他们的竞争力非常强，他们的业绩长期稳定增长，资产会越来越值钱。

又如，三峡能源的绿电供给很多的AI数据中心，收入长期稳定增长，资产也越来越值钱。

长期机会是壁垒最深的，也是最稳定的，因为这些资产是稀缺的，是不可复制的，一旦拥有了，就能够长期赚钱、永久赚钱。

21.4 终局：词元出海与垂直行业头部，利润上限的终极机会

终局，也就是10年以后，最确定的机会是词元出海与垂直行业头部。因为当所有的基础设施都建好之后，最终的利润会集中在最上游的词元生产者，也就是出海的词元厂商和垂直行业的头部玩家。这些是利润上限的终极机会。

词元出海，我们之前讲过，中国的词元在全球市场有碾压级的优势。未来全球的词元市场相关收入会达到3万亿美元，而中国的词元出口收入会占到30%，也就是9000亿美元。这是一个巨大的市场，而头部的出海厂商会垄断这个市场，拿到最大的利润。

比如，MiniMax现在已经是词元出海的龙头企业，其海外收入已经占到了70%，未来MiniMax会垄断全球的通用词元市场，年收入会超过1000亿美元，利润会超过500亿美元，这就是终极的机会。

还有垂直行业的头部玩家。每个垂直行业最终都会有一个头部的词元厂商，垄断这个行业的词元市场，拿到这个行业最大的利润。比如，医疗、法律、工业、金融的头部厂商，这些厂商会垄断整个行业，拿到永久的现金流。

终局的机会是利润最高的，也是壁垒最高的。因为这些厂商已经建立了自己的壁垒，锁定了用户，锁定了行业的标准，别人根本无法进入，他们会拿到整个行业最大的利润。

到这里，我们已经把词元经济的未来趋势全部讲完了。从词元成为基础能源，到全球产业链的重构，到下一代的词元工厂，到个人与企业的终极财富形态，到最后投资与布局的优先级，相信大家已经知道了词元经济的未来会是什么样子，也知道了该如何抓住这股浪潮带来的机遇。

词元经济的浪潮，是人类历史上最大的财富浪潮，它会彻底地改变整个世界经济、产业、财富的逻辑，它会创造比互联网多10倍甚至100倍的机会。希望大家能够在这个新的时代站稳脚跟，拥有属于自己的终极财富。

结语

在词元时代，发现属于你的新财富

到这里，本书的内容就已经全部呈现完毕，我们从词元的基础定义，到三层产业格局、“铁三角”的运作逻辑、财富的增长路径、生产线的搭建方法、商业化的落地路径，再到未来的长期趋势，我们将词元经济的所有核心内容，都系统、完整地梳理到位。

我知道，大家现在可能还有很多的疑问，可能会觉得这些内容是不是太遥远了，是不是太理想化了。

但是我要告诉你，这些都不是遥远的未来，这些都是现在正在发生的事情，我们在这本书里提到的所有案例、方法、路径，都是已经有案例验证过的，都是已经有人做到了的。

这些都不是未来的故事，这些都是现在正在真实发生的事情，这些都证明了我们在这本书里讲到的所有东西，都是真实的、可行的，都是能够做到的。

现在，你已经看完了这本书，已经知道了词元经济是什么，已经知道了怎么搭建自己的生产线，已经知道了怎么把生产线变成“印钞机”，已经知道了未来的趋势是什么，已经知道了怎样抓住这个机遇。

接下来，你要做的，就是行动。

很多人看完了书，知道了所有的道理，但还是不会行动，他们会说，再等等，再看看，等技术再成熟一点，等市场再稳定一点，然后他们就错过了大好的机遇。

在互联网时代，很多人是这么说的，他们等啊等，等到互联网成熟了，他们才入场，然后就错过了所有的红利。

在移动互联网时代，很多人也是这么说的，他们等啊等，等到移动互联网成熟了，他们才入场，然后也错过了所有的红利。

现在是词元经济的时代，现在就是最好的入场时机，现在入场，就能够吃到最大的红利，如果再等，就会像过去的很多次一样，错过了这个时代的最大机遇。

所有伟大的事业，都是从小开始的；所有的巨头，都是从小玩家成长起来的。MiniMax一开始也只是一个几人的小团队，创研股份一开始也只是一个普通的SaaS公司，但是他们行动了，他们抓住了机会，所以他们成功了。

现在你也可以，只要行动起来，也能够做到。

现在搭建生产线的门槛非常低，零代码的工具已经非常成熟，只要点几下鼠标，就能够搭建自己的生产线。

资源也不是问题，模型市场已经非常成熟了，OpenRouter已经把全球的用户都准备好了。只要把词元接上去，就能够拿到全球的订单。不需要自己做推广，不需要自己做交付，不需要自己做结算，所有的事情OpenRouter都已经帮你做好了。

你唯一需要的，就是行动，就是开始，就是迈出你的第一步。

因为词元经济时代已经到来，未来词元会像电力、石油一样，成为整个社会的基础能源，所有的智能活动、生产经营、生活消费，都离不开词元，词元会成为新的硬通货，词元产能会成为新的终极财富。

谁能够提前布局，提前拥有词元的产能，谁就能够在这个新的时代拥有最大的财富，就能够拥有永久的现金流，就能够拥有永久的竞争力。

黄仁勋在GTC的演讲里说，未来10年AI会彻底改变整个世界，所有的行业、所有的企业、所有的个人，都会被AI重构，而词元就是这一切的核心和基础。

这是人类历史上最大的技术革命和财富浪潮，它会创造出比过去所有时代加起来还要多的财富，它会给每一个愿意行动的人彻底改变自己命运的机会。

我希望，很多年以后，当你回头看时，你会感谢今天的你，感谢自己在这个最好的时机选择了行动，选择了抓住这个机会。你会庆幸，没有像

其他人一样，错过这个时代的红利。

这就是词元经济给你的机会，这就是AI时代给你的机会。

最后我想说，词元经济的时代已经到来，愿每一位读到这本书的人，能够抓住这个时代的机遇，实现财富自由，能够在这个新的时代，拥有属于自己的精彩人生。

词元经济：AI时代财富增长密码

作者 蒋凡

特约编辑 周延

装帧设计 杨杨

技术编辑 丁占旭

印务 杨景依

出品人 曹俊然

果麦

www.goldmye.com

以微小的力量推动文明