“龙虾”太能吃 钱包“顶不住”

这份“省元”攻略让你不花冤枉“元”

近日，国家数据局有关负责人在国务院新闻办新闻发布会上表示，到今年3月，我国日均词元（Token）的调用量已经超过了140万亿。随着AI视频创作的爆发、OpenClaw部署热潮的到来，词元调用量出现指数级增长，“如何减少词元的消耗”也成了许多大模型用户的热议话题。为此，深圳特区报邀请到了业内专家从实操技巧、输出管控等方面，分享减少词元消耗的方法，让你“养龙虾”不花冤枉钱。

词元 大模型处理信息的最小信息单元。词元巨大调用量的背后，是使用费用的快速增长。

词元为什么消耗得这么快？

词元，大模型处理信息的最小信息单元。假如你问ChatGPT：“今天天气怎么样？”这句话可能会被切成5个词元——“今天”是1个词元，“天气”是1个词元，“怎么样”可能被切成2个词元，句末的问号可能又是1个词元。

在AI眼里，不论是文字、音频、图片还是视频都是一个个词元。同样的，AI输出的答案，在它的标准里，也是一堆大小不等的词元。

伴随着AI视频创作的爆发、OpenClaw部署热潮的到来，词元调用量出现指数级增长。原因在于，此前用户与大模型的交互方式多局限于一问一答，只要用户停止回复，那么任务也随之终止。而OpenClaw的核心差异在于，它接收指令后会自动拆解任务并执行全流程——从搜集资料、编写代码，到调试程序、优化方案，每一步都需与大模型完成多轮交互，词元消耗随之倍增。在不久前举行的国务院新闻办新闻发布会上，国家数据局有关负责人表示，到今年3月，我国日均词元的调用量已经超过了140万亿，相比2024年初的1000亿增长了1000多倍，相比2025年底的100万亿，3个月的时间又增长了40%。

而词元巨大调用量的背后，是使用费用的快速增长。打开社交平台，词元消耗“大爆炸”的讨论密集涌来。一位中国开发者在阿里云开发者社区分享了自己的经历：使用OpenClaw进行自动化任务处理，2个小时就花费了100美元的词元消耗费。

初创企业AlayaDB架构师、南方科技大学和香港理工大学博士生游正新表示，词元的计费公式为输入词元数量×输入单价+输出词元数量×输出单价+其他特殊费用。同时，大模型采用的是“非对称计费”，即输入词元、输出词元的单价并不一样，“通常输入价要比输出价低很多，这也很符合我们的认知——输出内容要比理解观点困难得多。”

这些技巧帮助你精准“控本”

近期，第一批“养龙虾”的用户开始发现，这个AI牛马比想象中“贵”太多了，“如何减少词元消耗”自然也成了热议话题。作为词元消耗大户，游正新透露，自身每月词元调用量约600万至800万，仅购买各大主流模型API调用套餐，月均花费就达500元至1000元。

为此，他总结出了自己的一套减少词元消耗的方法：固化流程减少探索，将成功完成的复杂任务转化为可复用的skill和工具，后续可改用小模型驱动，替代高成本大模型；复用上下文，同一主题或文件的对话集中在同一个对话框进行，避免大模型需要重复学习而导致词元的大量消耗；提前替大模型规划好行动计划，减少它的额外思考甚至“走弯路”；清晰地表达需求，必要时让大模型优化提示词；给大模型明确的输出要求，减少整体词元的消耗。

“同时，对文字、图片、音频、视频进行预处理，也能减少AI在读取时的词元消耗。”游正新表示，对于文字输入，可精简指令与拆分对话，去掉礼貌用语、仅保留结构性指令；图片输入时预先裁切空白画面，不使用过高分辨率的图片；对于音频，可以剔除空白片段缩短时长，或者优先转成文字后再提交；视频则主打极简采样，将采样频率调至每5秒1帧，或提取关键帧以图片组形式提交，“这些小细节的处理，都可以大大减少词元的消耗。”他说。

值得关注的是，深圳正通过多重政策进一步降低用户词元使用成本：罗湖区对经认定的人工智能OPC，提供年度最高100万元的算力与模型调用支持；前海OPC国际社区则为入驻企业提供“零成本”算力福利——每年最高50P算力免费服务，同时开放主流大模型免费试用通道。与此同时，全市还推出了“模型券”“智能券”，可直接抵扣词元消耗费用。

（据深圳特区报）