AIGC 行业跟踪：CHATGPT 使用成本下降或将是打开产业应用市场的拐点

此次 OpenAI 使用成本的大幅下降，很有可能来自于模型架构的调整。3 月 2 日，OpenAI 在官方博客宣布，此次 OpenAI 开放的 ChatGPTAPI 模型是 Gpt-3.5-turbo。这与 ChatGPT 目前使用的是同一种模型。价格为 0.002 美元/1ktokens，GPT-3.5 模型达芬奇版本为 0.02 美元/1ktokens，使用成本下降 90%。

回溯 GPU 的发展历程，1）人工智能算力核心来源于 GPU，2010 年英伟达发布的 Fermi 架构，是第一个完整的 GPU 架构。其计算核心由 16 个 SM(StreamMultiprocesser)组成，每个 SM 包含 2 个线程束(Warp)，16 组加载存储单元（LD/ST）和 4 个特殊函数单元（SFU）组成。2）2016 年的 Pascal 架构，英伟达开始往深度学习方向演进。在 SM 内部，除了以往支持单精度的 FP32CudaCore，还增加了支持双精度的 DPUnit，而 DPUnit 实际上是 FP64 的 CudaCore。

3）从 CudaCore 到 TensorCore，通过精简业务模块，满足低精度输出要求，进而节省成本。2017 年以后，引入了张量核 TensorCore 模块，用于执行融合乘法加法。其中两个 4*4FP16 矩阵相乘，然后将结果添加到 4*4FP16 或 FP32 矩阵中，最终输出新的 4*4FP16 或 FP32 矩阵。我们认为此次 OpenAI 成本的大幅下降，很有可能来自于模型架构的调整。

ChatGPT 成本仍有降低空间，我们认为 ChatGPT 使用成本的下降或将是打开产业应用市场的拐点。1）我们认为此次成本的下降可能来自于对算法算力以及 GPU 的优化。包括业务层的优化，降低延迟和重复调用；模

型层优化，去掉作用不大的结构等等；量化优化，kernel 层优化，编译器层优等等。2）主导的变化：让以前高精度 CUDACore 为主要的运算降低到可以以 TensorCore 为主要的模型去跑，这样的话就可以大幅降低使用成本。3）往后看我们认为这种优化会持续不断进行，即使现在成本降低 90%，但是随着 OpenAI 技术的迅速迭代，展望未来还有进一步下降的空间，目前这个阶段可以认为已经看到初步应用的拐点。

建议关注：OpenAI 产业进展、研发进展提速，我们认为相关模型比国内领先，海外更快做好准备；成本优化成功，和搜索引擎在数量级上可以进行初步比较，我们认为海外产业链这个时点有充分逻辑，可以撬动下游需求和生态发展，所以海外业务占比越高的企业，撬动力更高。1）出海产业链（海外业务占比高）：

推荐福昕软件、建议关注昆仑万维、万兴科技。ChatGPT 降价对于行业未来发展有提振作用，对大规模商用有促进作用，对 AI 产业链拉动作用强。2）AI 大模型：建议关注 360、拓尔思、推荐科大讯飞；3）AI 供应链：建议关注海天瑞声、中兴通讯、推荐中科曙光、海光信息；4）英伟达产业