关于这几天很火的 DeepSeek，我们（Archerman Capital）做了一些研究和思考，和大家分享，enjoy！灰色部分是技术细节，不感兴趣的可略过。

几个事实

1）DeepSeek 不是套壳不是蒸馏美国的大模型。虽然中国有些大模型是套壳和蒸馏的，但 DeepSeek 不是。

2）核心架构还是基于 Transformer，deepseek 在架构、工程设计上进行了创新和工艺提升，实现效率优化。架构上，采用了混合专家模型 (MoE)、多头潜注意力 (MLA) 、多令牌预测（MTP）、长链式推理（CoT）、DualPipe 算法等设计，并进行了依赖强化学习（RL）而不加入监督微调（SFT）的训练尝试。工程上，在数据精度（FP8 混合精度）、底层通信等方面进行了优化。这些方法在学术界都已经有了，Deepseek 没有过于追求新技术，而是花了心思把这些方法都用上，解决了一些技术的应用难点，在理论应用和工程上找到平衡，具体如下：

- MoE：Mixture of Experts。将模型划分多个专家模块来进行分工。训练中将不同专家模块分配到不同计算设备训练，提升训练效率。推理时，仅动态激活部分专家（37B 参数），而非全模型参数（671B 参数），减少计算负担。但是 MoE 经常会面临某些专家承担所有工作，其他专家不被使用的问题，业内会通过一如辅助损失来对此调控、平衡各个专家模块的工作量，而 deepseek 通过无辅助损失的自然负载均衡（引入一个无形的手而不是人为调控）、共享专家机制来解决该问题。
- MLA：Multi-Head Latent Attention。扩展了传统的多头注意力机制，引入潜向量（latent variables），可以动态调整注意力机制，捕捉任务中不同的隐含语义。在训练中减少内存和计算开销，在推理中降低 KV 缓存占用空间。
- MTP：Multi-Token Prediction。一般 LLM 一次生成 1 个 token，采用单步预测。deepseek 在特定场景下能同时预测多个 token，来提高信号密度。一方面能够减少上下文漂移、逻辑更连贯，也能减少一些重复中间步骤，在数学、代码和文本摘要场景能提升效率。
- Cot：Chain of thought。一种训练和推理方法，将复杂的问题拆分成小步的中间逻辑，细分逻辑链条。在训练阶段，Deepseek 用标注的 Long CoT 数据微调模型，让模型生成更清晰的推理步骤，在强化学习中用 CoT 设计奖励优化，增强长链推理能力，并且在此过程中观察到了模型的反思（回溯推理路径）、多路径推理（能给出多个解）、aha 时刻（通过策略突破瓶颈）等自发行为。

- DualPipe：传统训练信息流水线会产生一些等待时间、有“流水线气泡”，deepseek 设计了一个双重流水线，让一个计算阶段在等待数据传输时可以切换到另一批数据，充分利用空闲时间。
- R1-Zero：Deepseek 在 V3 基础模型上，仅通过强化学习（Rreinforcement Learning）训练，而不加入 SFT（Supervised fine tuning）数据，训练了 R1-Zero 模型，探索了模型不依赖人类标注数据微调、自主推演的能力，打开了新的思路。但 R1 模型仍然采取 SFT 数据优化推理和生成质量。
- FP8 混合精度训练：引入了 FP8 混合精度训练框架，相比传统的 FP16 精度，数据内存占用更少，但在一些算子模块、权重中仍然保留了 FP16、FP32 的精度，节省计算资源。
- 底层通信优化：开发了高效的通信内核，优化对带宽的利用，保证数据传输效率，并能支持大规模部署。

拿内燃机和汽车的发明打个比方，德国人发明了内燃机和汽车，美国人喜欢 Scaling Law，排量越大马力越大，于是从 2 升到 4 升，甚至 8 升排量的车在美国都很常见，所以美国肌肉车很耗油。虽然源头技术不是日本发明的，但日本人擅长把一件事做精，工程上做很多优化，日本 2.5 升排量的车甚至可以做到和美国 5 升排量车一样的百公里加速指标。比如轻量化设计把大钢板换成钢条（类似通过稀疏的办法减少大模型的参数量）；涡轮增压利用废气能量增加空气供给，提高燃烧效率；精密制造，使得发动机零部件的配合更加紧密，从而减少能量损失；等等。

3）有些宣传说 DeepSeek 的训练成本是 550 万美元，是 Meta 的 1/10，OpenAI 的 1/20，好像一下子比别人厉害了 10 倍 20 倍，这有点夸张。因为现在在美国预训练几千亿参数的一个模型其实也就不到 2000 万美元的成本，DeepSeek 把成本差不多压缩到三分之一。Meta 和 OpenAI 花的钱多是因为前沿探路，探路就意味着会有浪费，而后发追赶是站在别人的肩膀上，是可以避开很多浪费的。另外算力成本在过去几年是指数型下降的，不能这么机械的比较。打个不恰当的比方，创新药的研发需要十年几十亿美元，而仿制药的研发一定会更快更省。另外成本的统计口径也没有统一的标准，可以有很大的差别。

几个观点：

1）DeepSeek 代表的是整个开源相对闭源的一次胜利，对社区的贡献会快速转化为整个开源社区的繁荣，我相信包括 Meta 在内的开源力量，会在此基础上进一步发展开源模型，开源就是一个众人拾

柴火焰高的事情。

2) OpenAI 这种大力出奇迹的路径暂时看显得有点简单粗暴，但也不排除到了一定的量又出现了新的质变，那闭源和开源又将拉开差距，这也不好说。从 AI 过去 70 年发展的历史经验来看算力至关重要，未来可能依然是。

3) DeepSeek 让开源模型和闭源模型一样好，并且效率还更高，花钱买 OpenAI 的 API 的必要性降低了，私有部署和自主微调会为下游应用提供更大的发展空间，未来一两年，大概率将见证更丰富的推理芯片产品，更繁荣的 LLM 应用生态。

4) 基础大模型终将 commoditize（商品化），toB 领域看谁能将 LLM 更好和复杂的生产环节衔接好帮客户落地提高生产效率，toC 领域看谁有流量入口，最终才会获取 AI 产业价值创造中最多的利润。

5) 对算力的需求不会下降，有个 Jevons 悖论讲的是第一次工业革命期间蒸汽机效率的提高使得市场上煤炭的消耗总量反而增加了。类似从大哥大年代到诺基亚手机普及的年代，正因为便宜了所以才能普及，因为普及了所以市场总消费量增加了的。

6) 对数据的需求不会降低，巧妇难成无米之炊，没有米怎么做饭，算法的提高相当于做饭吃饭变得更快，对数据的渴求会更大。

研究期间，我们与几位学术界和工业界的专家进行了交流，由于尚未获得公开提名的许可，就暂不提及具体姓名了，但在此特别表达感谢！ Archerman Capital™ 是一家美国的成长期股权投资机构，专注于人工智能、数据基础设施、网络安全等领域的成长期投资。其投资组合包括 Databricks, Scale AI, Tenstorrent 等。该机构采用高度研究驱动和第一性原理的方法。公司总部位于波士顿，在纽约和硅谷设有投资团队。以上是纯分享，并非投资建议。