亮点段落

- 该研究表明，当前大语言模型（LLM）的迅速扩展暴露了现有硬件架构的许多局限性，比如内存容量、计
  V3在2048块 NVIDIA H800GPU 集群上训练，通过有效的硬件感知模型设计，克服了这些限制，实现了
- 论文中提出了几个关键点。首先，DeepSeek-V3采用了先进的 DeepSeek-MoE 架构和多头潜在注意力（MLA）架构，极大地提高了内存效率。MLA 技术通过压缩键值缓存，显著降低了内存使用，使得每个 token 只需70KB 的内存，相比其他模型大幅减少。
- 其次，DeepSeek 还实现了成本效益的优化。通过其混合专家（MoE）架构，DeepSeek-V3在激活参数的数量上实现了显著的降低，训练成本相比于传统密集模型降低了一个数量级。此外，该
  最大 化吞吐量，确保 GPU 资源得到充分利用。

Figure 1: image.png|Left|700

作者通过结合硬件感知的模型设计与创新架构，有效突破了大规模训练的瓶颈，体现了技术与工程的深度融合

总结：在大规模AI模型的训练与推理中，硬件与模型架构的协同优化是突破性能瓶颈的关键所在。