正规分分彩app下载

EMCm7DuGMf9lBRLV

正规分分彩app下载人人都能有一个满血DeepSeek？清华90后出手，10万块畅玩FP8原版

新智元报道

编辑：编辑部 HYNZ

【新智元导读】最近，一款由清华90后团队打造的AI利器，首次实现了只用14.9万元就能流畅运行满血版DeepSeek，并且还支持顶配的128K上下文，堪称媲美百万级方案。

DeepSeek-R1从开源到现在，已过去4个多月。

然而，高昂的硬件成本、复杂的私有化部署方案，长期以来如同天堑，隔断了一些企业、开发者们拥抱AI的步伐。

但如今，人手一个满血版DeepSeek-R1/V3 671B的时代来了！

说出来你可能不信，行云褐蚁一体机把曾经上百万才能拥有的体验，直接打到了14.9万元。

不仅如此，它的速度和精度更是堪比官方版本——能以超过20 token/s的速度，运行没有经过量化的「FP8原版」DeepSeek模型。

这是什么概念？请看对比。

画面右侧就是DeepSeek官网的输出速度。而左侧，便是我们部署在一体机上的满血版DeepSeek-R1了。

原视频加速2倍

值得一提的是，刚刚发布的Qwen3超大杯MoE——235B-A22B，现在也可以用上了！

打造出性价比如此惊人产品的公司，是什么来头？

原来，这家的CEO正是前华为天才少年、清华90后博士季宇。

10万价位，单机可跑满血DeepSeek

接下来，我们在真机上实际测试一下，这个「原版」DeepSeek-R1到底怎么样。

先让它推理出一个笑话的笑点在哪里。

可以看出这个速度刷刷的，等待时间也很短，几乎可以忽略不计。

稍微长点的提示词，对速度也有明显的影响。

这道题只给出了一些疑似有规律的符号，而满血DeepSeek-R1则展示出了强大的推理能力，猜出这是词牌格式《菩萨蛮》，甚至猜出了是哪些符号分别对应的平、仄。

来一个甘蔗过门这种极易迷惑模型的问题。

DeepSeek-R1顺利给出了正确答案，速度也无可挑剔。

甚至，它非常顺利地做出了一道AIME 2025数学竞赛题。

即使是这种级别的推理，输出速度也能保持在20 token/s。

外星人来到地球上，可能会选择四件事中的一件来完成，求地球上最终没有外星人的概率。这种复杂的数学推理题，DeepSeek-R1也顺利做了出来。

因为可以在Dify工作流中使用，这台一体机甚至能完成DeepResearch的功能。

由于模型部署在本地，所以可以基于内部的私域数据进行深度挖掘和研究服务，保障信息隐私及安全。

广泛适用于文档摘要、数据分析、代码生成等高精度复杂任务。

完整工作流如下：

极致性价比

为什么褐蚁一体机，能用10万元的水平，达到以上惊人的水准？

背后原因，除了得益于自研的高效推理引擎外，还有极致的硬件选型。

褐蚁系列一体机有三种型号可选，理论上参数在1.5T以内的模型都能支持。

甚至，即将推出的DeepSeek-R2，预计也可以实现支持。

其中HY90负责提供极致的性能，671B参数的满血DeepSeek-R1/V3在最高精度FP8下，速度能达到21.5+ token/s；在FP4精度下，速度能达到28+ token/s。

FP8

INT4

HY70提供极致的性价比，同样是满血FP8精度的DeepSeek-R1/V3，速度也能达到20+ token/s，在FP4精度下，速度能达到24+ token/s，相当炸裂。

最后，HY50还提供了极致低价。支持671B参数的满血DeepSeek-R1/V3，在INT4精度下可实现20+ token/s的输出速度，相当实用。

不仅如此，褐蚁系列一体机支持多种AI推理引擎，支持API调用、知识库、AI Agent部署，还支持全部的开源大模型。

一次购买，永久使用。

在技术实现上，行云团队尽可能提高了大模型推理时有效使用的带宽上限（理论带宽1200 GB/s，物理实测1050 GB/s）。

而在实际使用中，这套系统的等效带宽可以达到800 GB/s，完全满足740 GB/s的需求。

算力层面，团队则通过一套独家定制的软件协同优化方案，极大地提升了系统的运行效率。

输出方面，上下文长度对速度的影响被控制得很好。只有当长度达到32K以上时，才有一些明显的下降。

prefill方面，16k以内可以保持在180～200 token/s左右，上下文首字延迟则在80秒以内。

具体来说，首字延迟在1k下是5秒，4k是20秒，8k是40秒，16k是80秒。不过，在128K极限上下文长度下，会达到30分钟。

LLM端侧部署，CPU了解一下

说到模型的本地化部署，通常的第一反应就是GPU服务器。

的确，在大模型训练时，GPU的优势可谓是独步天下——吞吐量可以达到CPU的数十倍甚至上百倍。

然而，在利用模型进行推理的应用阶段，一个缺点就足以把众多企业挡在门外——太贵！

以FP8精度为例：

671B的参数量，意味着需要671GB以上的内存

37B的激活参数，对应的是37GB x 20 token/s = 740GB/s以上的内存带宽

也就是差不多一套6卡H20 141GB，或者10卡A100/A800 80GB服务器才能跑起来。

即使按照目前市场上比较便宜的报价，这套系统的最低也要百万元以上。

为了降低成本，一些企业会采用模型量化，甚至是牺牲对话速度，来降低LLM对硬件的需求。

然而，量化会显著降低模型精度，尤其是在法律、医疗等需要高质量输出的场景中，可能会造成生成的结果不可靠。

而降低对话速度，则会破坏实时交互体验，客户可能因响应过慢而逐渐流失。

这种体验与成本的权衡困境，使得许多企业陷入两难——要么投入巨资追求高质量部署，要么选择低成本方案但牺牲应用效果。

结果是，LLM应用场景被局限在少数高预算领域，难以在更广泛行业中实现落地规模化。

以上，这些痛点共同构成了LLM端侧部署的「不可能三角」：成本、性能、体验三者难以兼得。

既然传统的GPU解决方案无法做到，为什么不考虑换个思路呢？

相比于用大量GPU去堆叠显存，CPU的性价比就高得多了。

中高端服务器中CPU的单颗价格，通常只在数千美元，这就落在了很多企业的可承受范围之内。

而且，CPU一直以来最大的短板——内存带宽，如今也有了解决方案。

比如行云的褐蚁一体机，就通过双路AMD EPYC 9355 CPU，在24条频率高达6400MT/s的64GB内存加持下，实现1.5TB的容量和1.2TB/s的带宽。

不仅完美满足要求，甚至还有冗余。

值得注意的是，为了改善CPU在推理过程中存在的算力不足情况，此时还需加入一张中高端GPU作为补充。

更令人惊喜的是，10万的价位还可以压得更低！如果降低对TPS体验的需求或原版精度的需求，甚至可以压缩到5万。

清华90后创业，明星资本加持

在这款产品背后，是一支由清华90后领衔，兼具学术深度与行业实战经验的创始团队。

灵魂人物，便是创始人兼CEO季宇，是一位妥妥的「天才少年」。

他本科就读于清华物理系，随后转向计算机系，并获得了计算机体系结构（AI芯片方向）的博士学位。

博士毕业后，季宇入选了华为天才少年计划。

在学术方面，季宇的成就同样令人瞩目。

他长期专注于AI编译器优化和处理器微架构等前沿难题，积累了深厚的AI芯片经验。

而且，作为共同一作在顶刊Nature发表了计算机体系结构论文，荣获了计算机学会CCF优博奖。

这些经历，为他日后创立行云，打造低成本、高性能褐蚁一体机奠定了坚实基础。

CTO余洪敏则有着深厚的学术背景，以及丰富的行业经验。他毕业于华科大，后在中国科学院半导体研究所获得博士学位。

余洪敏同样有着堪称豪华的职业履历。

他不仅出任过多款顶尖国产芯片的负责人和研发总监，而且还长期领导和管理超100人研发团队，精通芯片研发设计全流程，成功知道了10+款芯片流片与量产。

他多次推动先进工艺数据中心芯片的架构设计、工程实现，以及大规模商用，积累了无可比拟的实战经验。

行云集成电路的吸引力，不仅体现在技术和团队上，还得到了资本市场的广泛认可。

去年11月，行云完成了新一轮数亿元融资，投资方包括智谱AI、中科创星、奇绩创坛、水木清华校友基金、嘉御资本、春华资本等一众明星资本。

从成立到融资，行云仅用了一年多的时间，就在AI芯片领域站稳了脚跟。

行云的崛起，正是「中国初创」加速赶超的缩影。

从模型竞赛，到应用为王

行云褐蚁一体机的横空出世，如同一记重拳，击碎了大模型部署高成本的壁垒。

它的推出，不仅是技术层面的突破，更是顺应了端侧部署的三大趋势。

首先，是成本门槛的指数级下降。

过去私有化部署的成本以百万计，行云直接将其拉低至10万，未来甚至可能降至5万。

这种成本的骤降，让中小企业，初创公司乃至个人开发者，都能负担起高性能AI解决方案，极大地拓展了AI应用的边界。

其次，CPU方案的崛起，让硬件架构更加多元化。

GPU因显存容量和互联成本的限制，在LLM部署中逐渐显露瓶颈。行云的CPU内存方案证明，服务器CPU高带宽和超大容量内存，能够以更低成本满足需求。

未来，更多芯片厂可能转向类似CPU主导或混合的架构，推动硬件方案的多元化。

最后，应用爆发，会推动AI普惠化的进程。显而易见的是，AI行业已从单纯的模型参数竞赛，转向应用落地的比拼。

行云低成本、高性能解决方案，为教育、医疗、零售等行业的AI应用打开了大门。

这不仅是一款产品的胜利，更是一个时代的开端——大模型正从少数巨头实验室珍宝，转变为千行百业的标配引擎。

普惠AI的时代，已然启幕！

目前， 行云 褐蚁一体机已开放预约体验， 详情可进入官方公众号咨询 。

澳洲10全天人工免费计划软件

澳洲幸运10开奖结果皇家软件

澳洲幸运5全国开奖官网

澳洲5 计划

澳洲幸运10漏洞公式

澳洲10在线计划

澳洲10开奖网168

168澳洲幸运10正规官网网址

168全国统一开奖官网

澳洲5历史开奖号码记录查询

2020澳洲幸运10开奖结果历史记录查询

澳洲幸运10官网免费下载

幸运飞行艇官网开奖结果app下载

众赢国际版zygjbcom

澳洲十全天计划网页版

168澳洲幸运10开奖计划

澳洲幸运5开奖结果历史

6码345678不死规律图片

辜运飞艇开奖记录