1、前提概要

Deepseek 7B 量级模型，以 DeepSeek-R1-7B-Distill 模型为例，模型推理性能超越 GPT4o，半精度运行所需显存为 4.7G，全精度运行所需显存为 14G。所需的主流显卡显存容量不少于 16G，RAM 建议至少为 32G，并开启 Resize BAR 技术，允许调整 PCIe 设备的完整可映射内存/寄存器地址访问。为保证 CPU 不拖累显卡的整体性能，建议最低 I5 12600KF/R7 5700X 级别的 CPU 起步，CPU 核心数量建议不少于 8 核。

2、具体显卡型号

NVIDIA：

40 系显卡：

GEFORCE RTX 4090 显存容量 24G (注意不带 D，D 锁 AI 算力)

GEFORCE RTX 4080 显存容量 16G

GEFORCE RTX 4080 SUPER 显存容量 16G

GEFORCE RTX 4070 Ti SUPER 显存容量 16G

50 系显卡：

GEFORCE RTX 5090 显存容量 32G(注意不带 D，D 锁 AI 算力)

GEFORCE RTX 5080 显存容量 16G

GEFORCE RTX 5070 TI 显存容量 16G

AMD：

7000 系显卡 (DeepSeek 牛逼，A 卡推理基本不输 N 卡，打破了 CUDA 的护城河，性价比还高的离谱)：

Radeon RX 7900 XTX 显存容量 24G

Radeon RX 7900 XT 显存容量 20G

Radeon RX 7900 GRE 显存容量 16G

Radeon RX 7800 XT 显存容量 16G

国产显卡：

摩尔线程 S80 (适配性还存在问题，谨慎考虑，且推理速度较慢)

华为 Ascend 昇腾 910B (生态较为完善，可以高速推理运行，但价格较为昂贵)

3、具体整机配置推荐

N 卡：

CPU： I5 12600KF 950 元

主板：铭瑄终结者 B760M 主板 800 元

内存：阿斯嘉特 32G DDR4 3600 MHZ 弗雷电竞条 350 元

固态 宏碁 GM7 2TB 600 元

显卡：RTX 4070Ti SUPER 6300 元

其它：机箱+电源+散热器+风扇等，共计 1000 元

总价：10000 元
A 卡：
配置同上，显卡更换为 RX 7800XT 3500 元
总价：7200 元

国产配置单：
无独立显卡龙芯 3A6000 整机，6999 元
显卡：摩尔线程 S80 1500 元
总价：8499 元（现阶段国产硬件总价性价比仍然有待进一步提高，但相较以前有很大提高了，未来可期）