Qwen2 7B模型调研

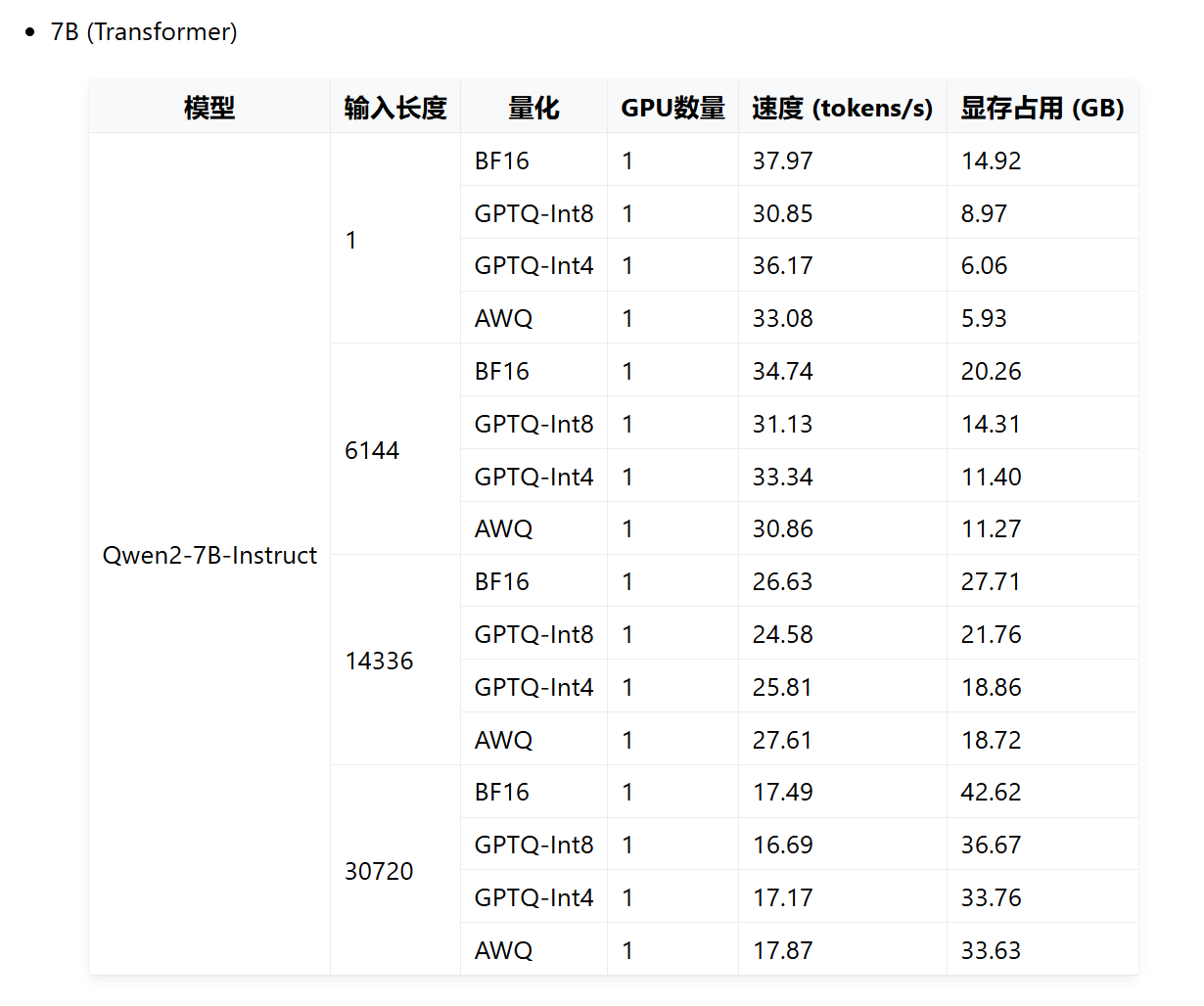
0. 配置要求

0.1 搜索收集

通过不同渠道搜集后，发现对于本地大模型的部署，大多数的文章只考虑参数在几十B以上的大模型，这些模型一般都需要服务器配置（多张3090以上显卡以及64GB以上显存），而对于本文讨论的7B模型，配置一般无需这么高。

下面是一些相关文章的分析：

[效率评估 - Qwen](https://qwen.readthedocs.io/zh-cn/latest/benchmark/speed_benchmark.html)



上图为官方文档截图，观察GPTQ-Int4参数，我们实际需要的显存在8-10GB左右？

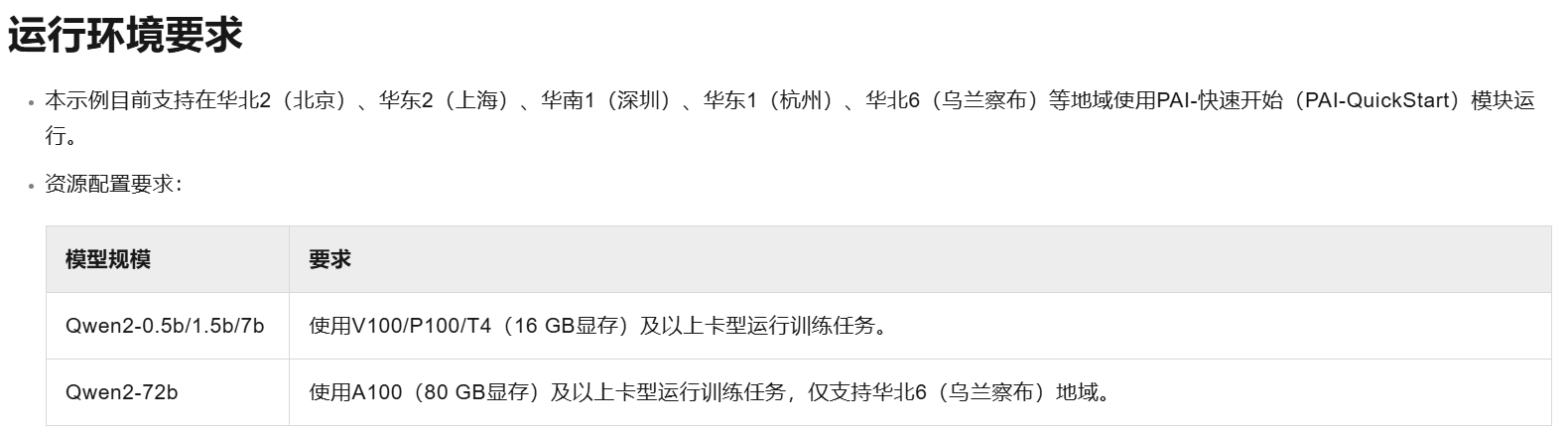
[Llama-2 推理和微调的硬件要求总结：RTX 3080 就可以微调最小模型 - 知乎](https://zhuanlan.zhihu.com/p/655948272)

根据这一篇Blog，如果需要微调大模型，需要至少RTX3060 以及6GB显存，为了更好的效果，估计需要至少8GB显存

[ollama本地部署大模型（纯CPU推理）实践\_ollama cpu-CSDN博客](https://blog.csdn.net/yang2330648064/article/details/137185603)

根据这一篇Blog，运行7B至少需要8 GB可用 RAM，运行13B型号至少需要16 GB可用 RAM，运行33B型号至少需要32 GB 可用 RAM。

[通义千问2模型部署与微调\_人工智能平台 PAI(PAI)-阿里云帮助中心 (aliyun.com)](https://help.aliyun.com/zh/pai/user-guide/deploy-fine-tune-and-evaluate-a-qwen2-model)



这篇Blog指出了我们使用的Qwen2 7B模型需要V100/P100/T4（16GB显存）及以上显卡。经搜索，上述卡均为专用计算卡，但是事实上由于经济原因并且我们并不需要这么高的算力，我们并不选择这些显卡。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Release Date | Max Length | System Prompt Enhancement | # of Pretrained Tokens | Minimum GPU Memory Usage of Finetuning (Q-Lora) | Minimum GPU Usage of Generating 2048 Tokens (Int4) | Tool Usage |
| Qwen-1.8B | 23.11.30 | 32K | ✅ | 2.2T | 5.8GB | 2.9GB | ✅ |
| Qwen-7B | 23.08.03 | 32K | ❎ | 2.4T | 11.5GB | 8.2GB | ✅ |
| Qwen-14B | 23.09.25 | 8K | ❎ | 3.0T | 18.7GB | 13.0GB | ✅ |
| Qwen-72B | 23.11.30 | 32K | ✅ | 3.0T | 61.4GB | 48.9GB | ✅ |

0.2 配置需求

综合上述文章，我们为了运行并且对Qwen2 7B模型进行训练，需要3060以上显卡以及8GB以上显存。



1. 能力边界

参考文章：

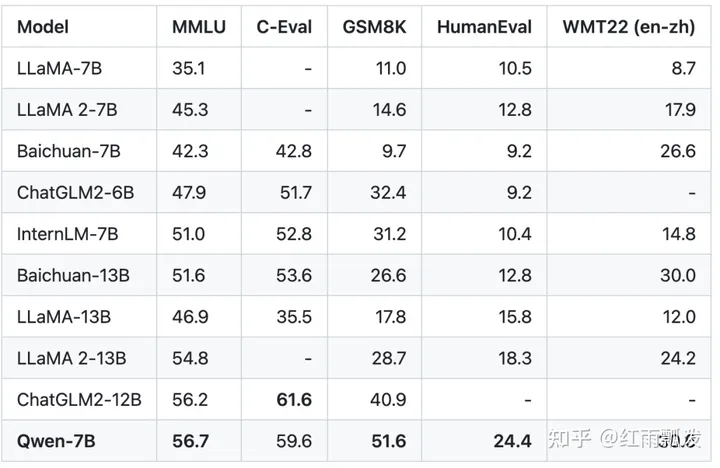
<https://www.zhihu.com/question/658347398/answer/3529444995>

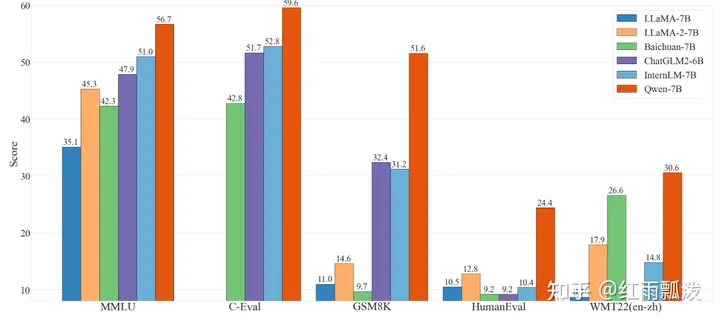
<https://zhuanlan.zhihu.com/p/702475296>

<https://qwenlm.github.io/zh/blog/qwen2/>

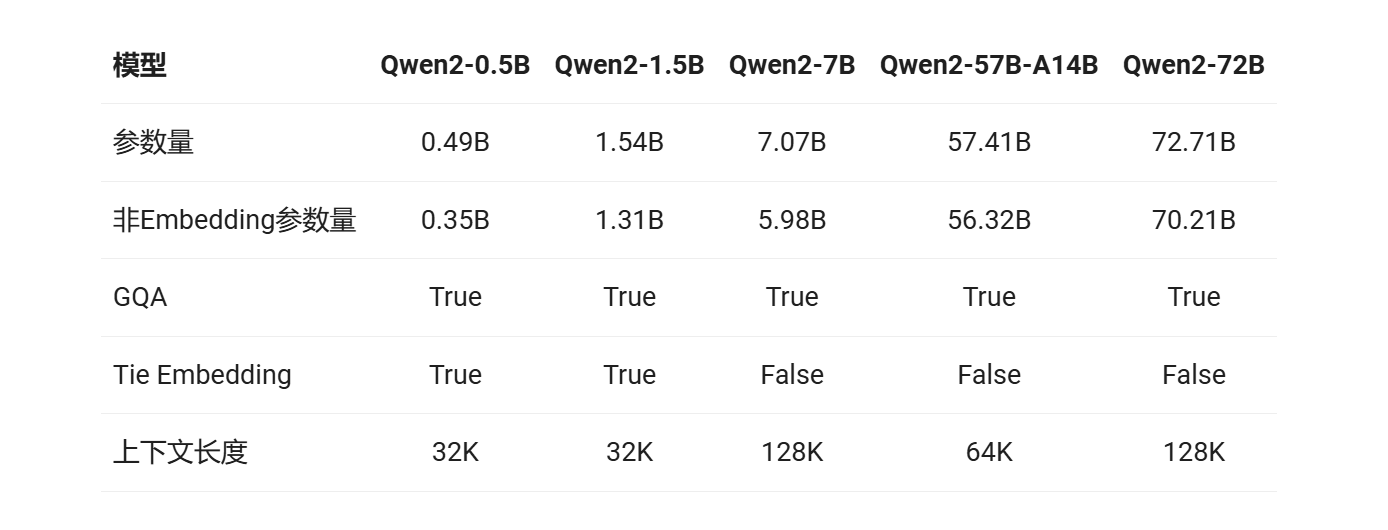
1.1 基本情况对比

Qwen2-7B在多个全面评估自然语言理解与生成、数学运算解题、代码生成等能力的评测数据集上，包括MMLU、C-Eval、GSM8K、HumanEval、WMT22等，均超出了同规模大语言模型的表现，甚至超出了如12-13B参数等更大规模的语言模型。





1.2 长文本阅读能力

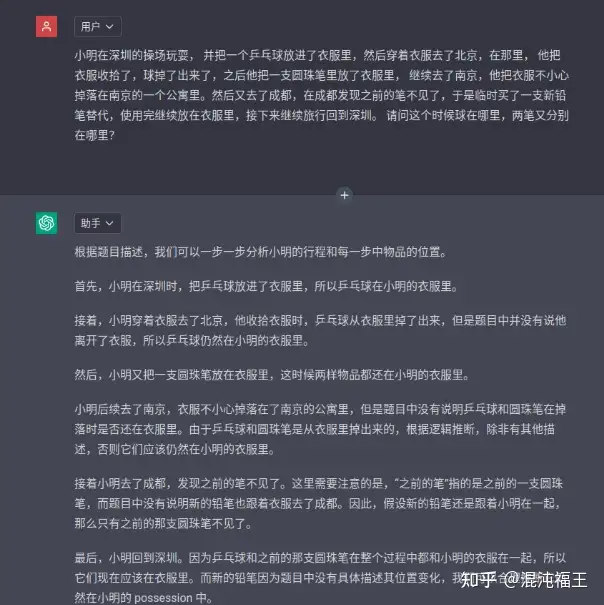


根据Qwen2 的官方文档， Qwen2 7B最长支持128K tokens，然而在需要阅读32k以上长上下文时，需要使用vLLM部署模型并启用长上下文功能才能取得较好功能（启用长上下文功能时可能会影响短文本性能）。

1.3 翻译能力

英译中时虽然也有可能出现中英文混杂的情况，但概率远低于14B模型，且翻译结果更加地道，显示出更好的翻译质量。

1.4 逻辑推理



对于复杂的逻辑问题，Qwen2-7b相较于其他参数量相同的大模型，要解决的更好

1.5 文件读取

对于文件读取，我们自己部署的大模型似乎在文件上传时有一些BUG，但是我们并没有找到解决方法，ModelScope官网提供了在线的Qwen2模型，但没有提供文件上上传入口。同时网络调研显示，似乎并没有相关文章显示文件读取的上限，但是，大模型读取类似.pdf之类的文本文件时，一般会调用Python的专用于文档读取的库，先将其转换为txt文件，再上传到大模型进行处理，所以Qwen2-7b的文本处理能力是可以对于该问题给予参考价值的。

1.6 代码能力



Qwen2-7b的代码能力在需求逻辑简单的情况下表现良好，但一旦需求逻辑变得复杂，其生成的代码逻辑会出现一定的混乱，功能会出现缺少，不可用等情况。