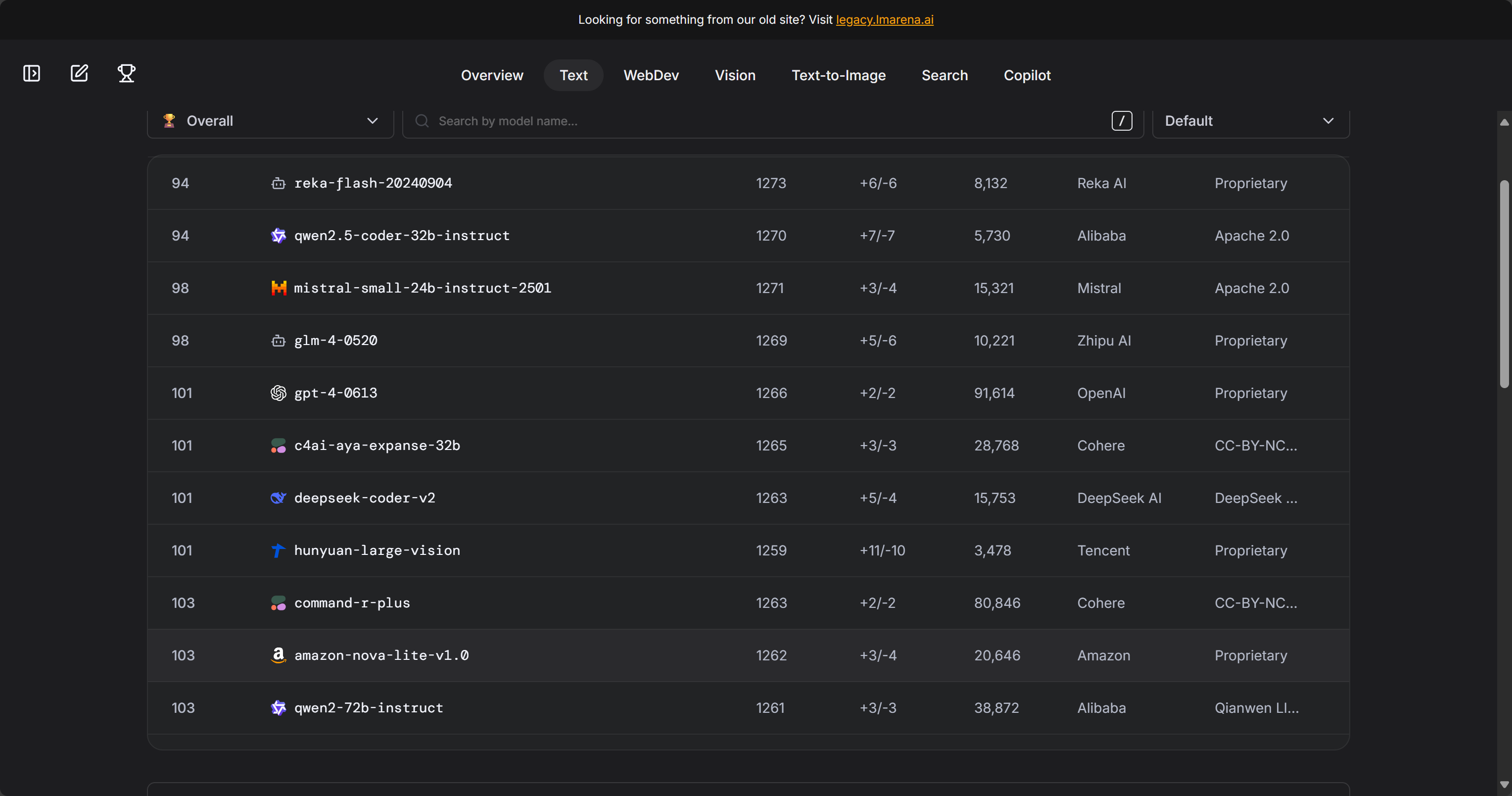
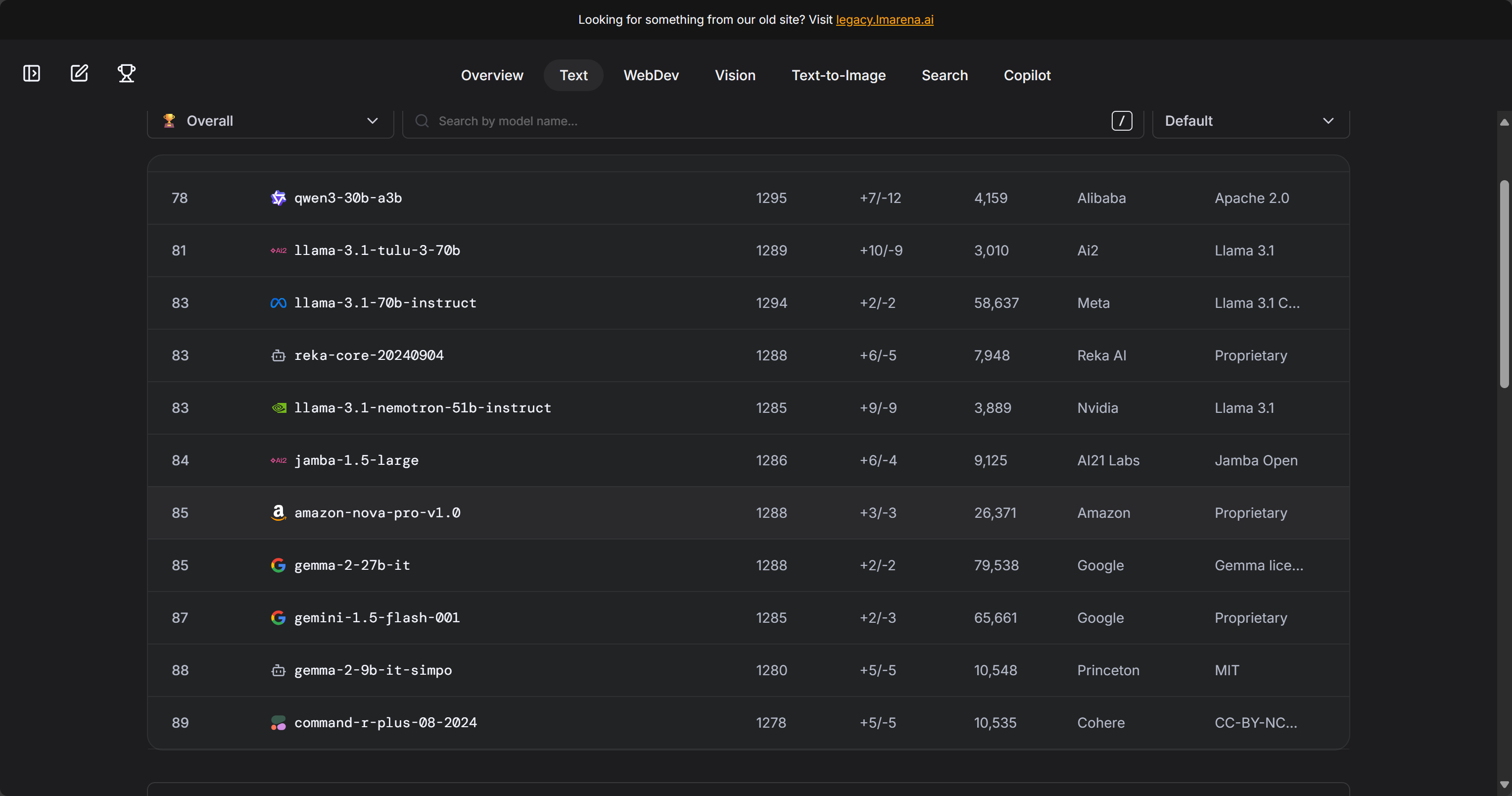
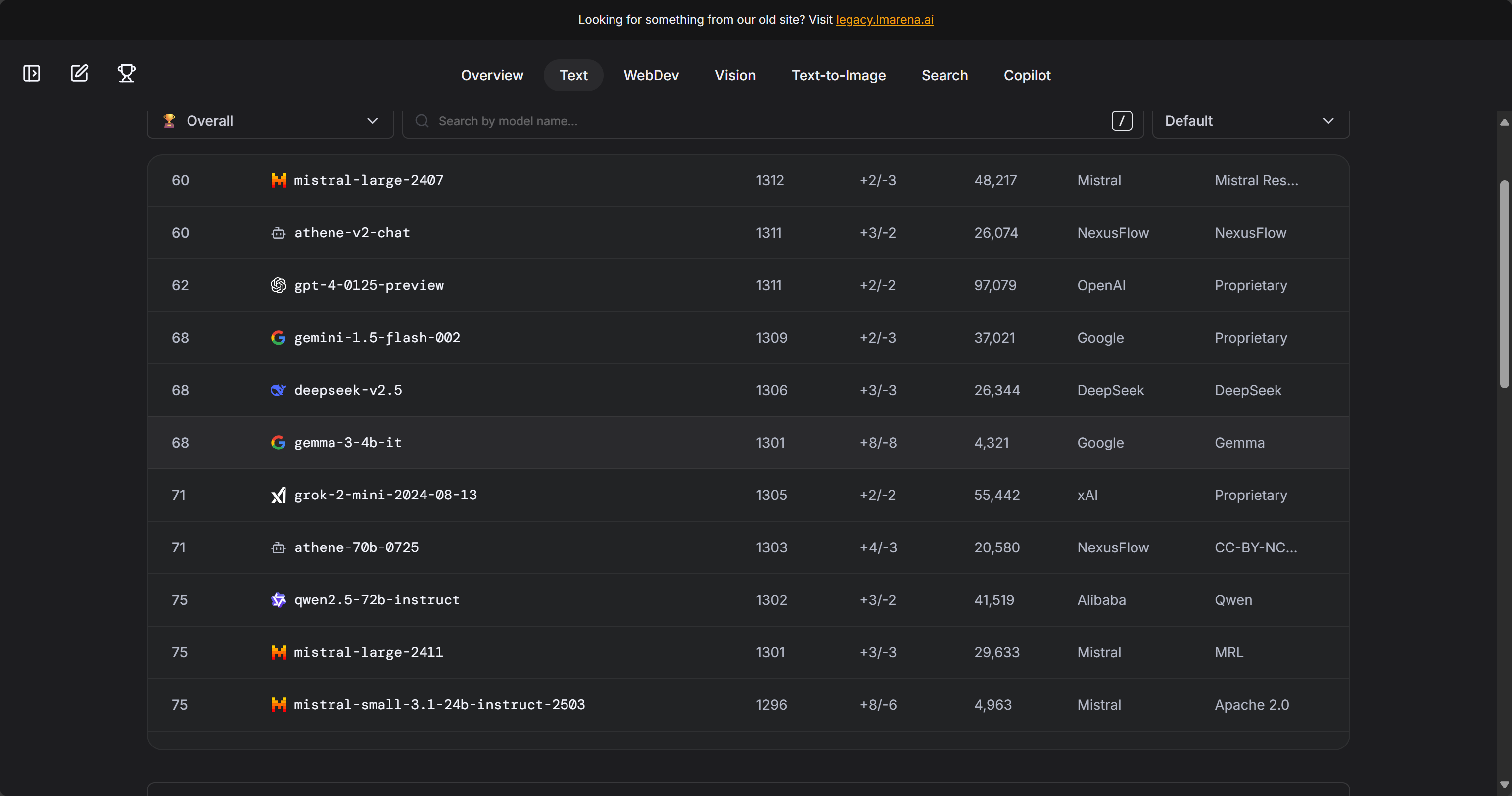
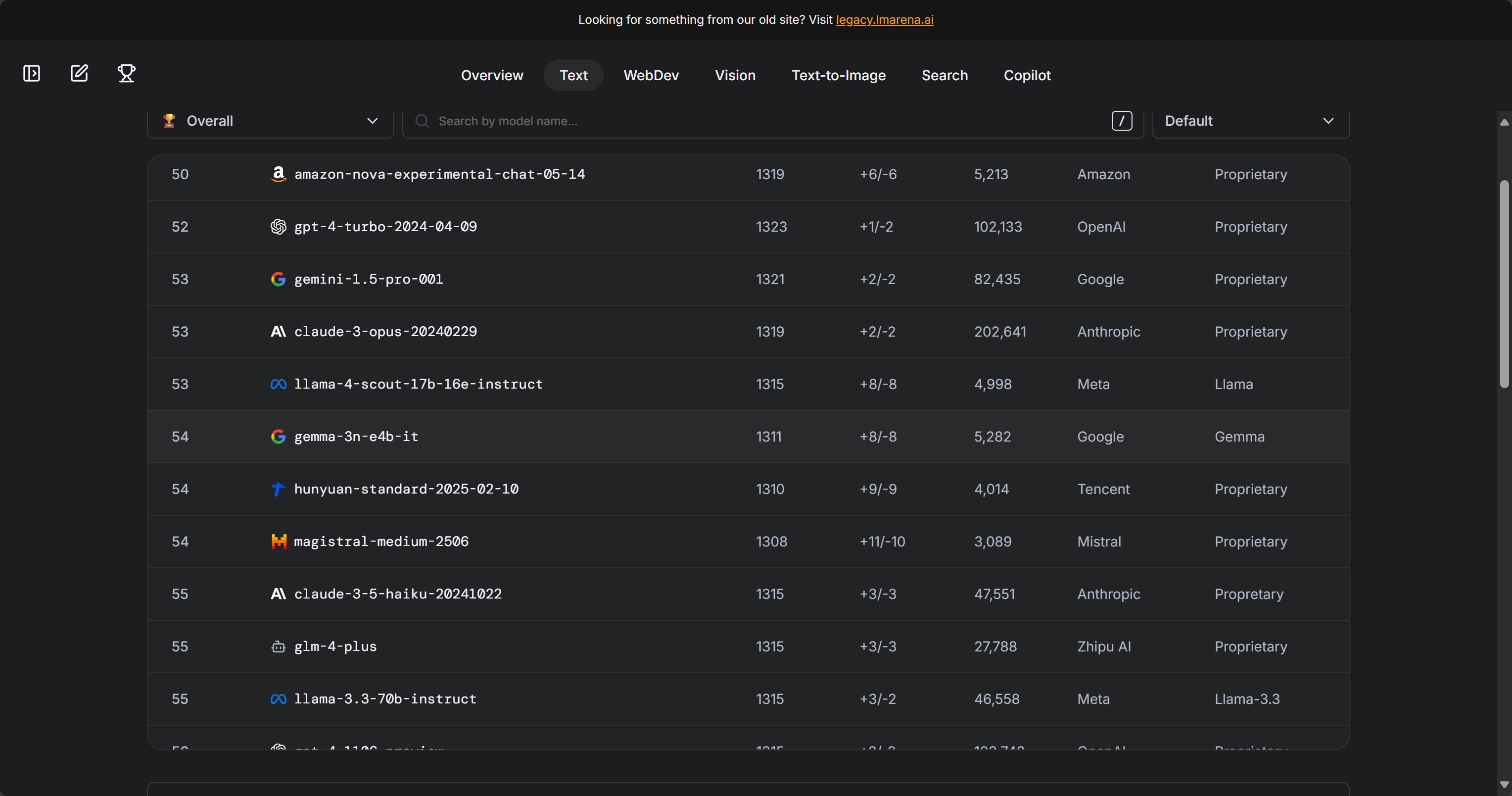
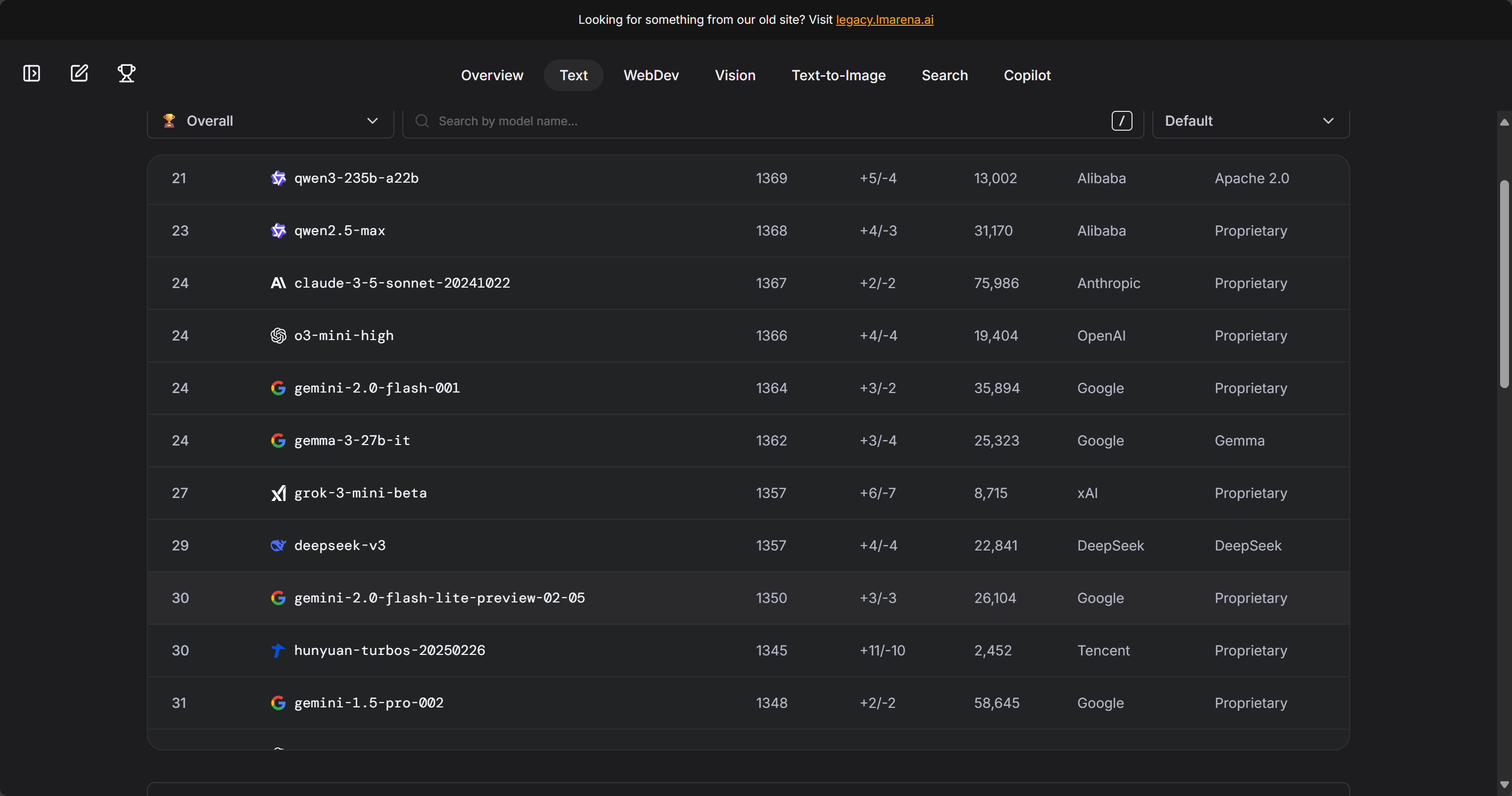
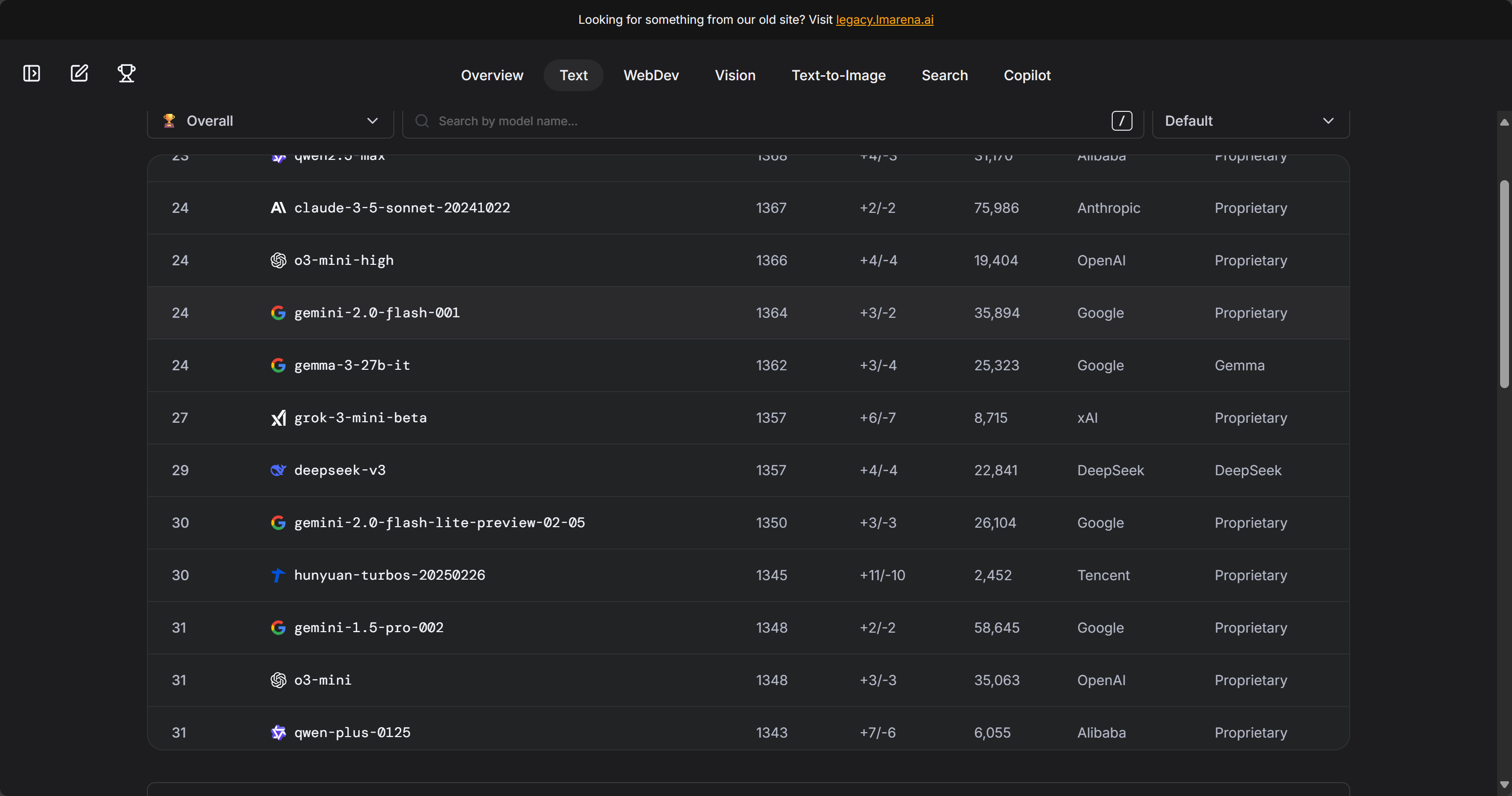
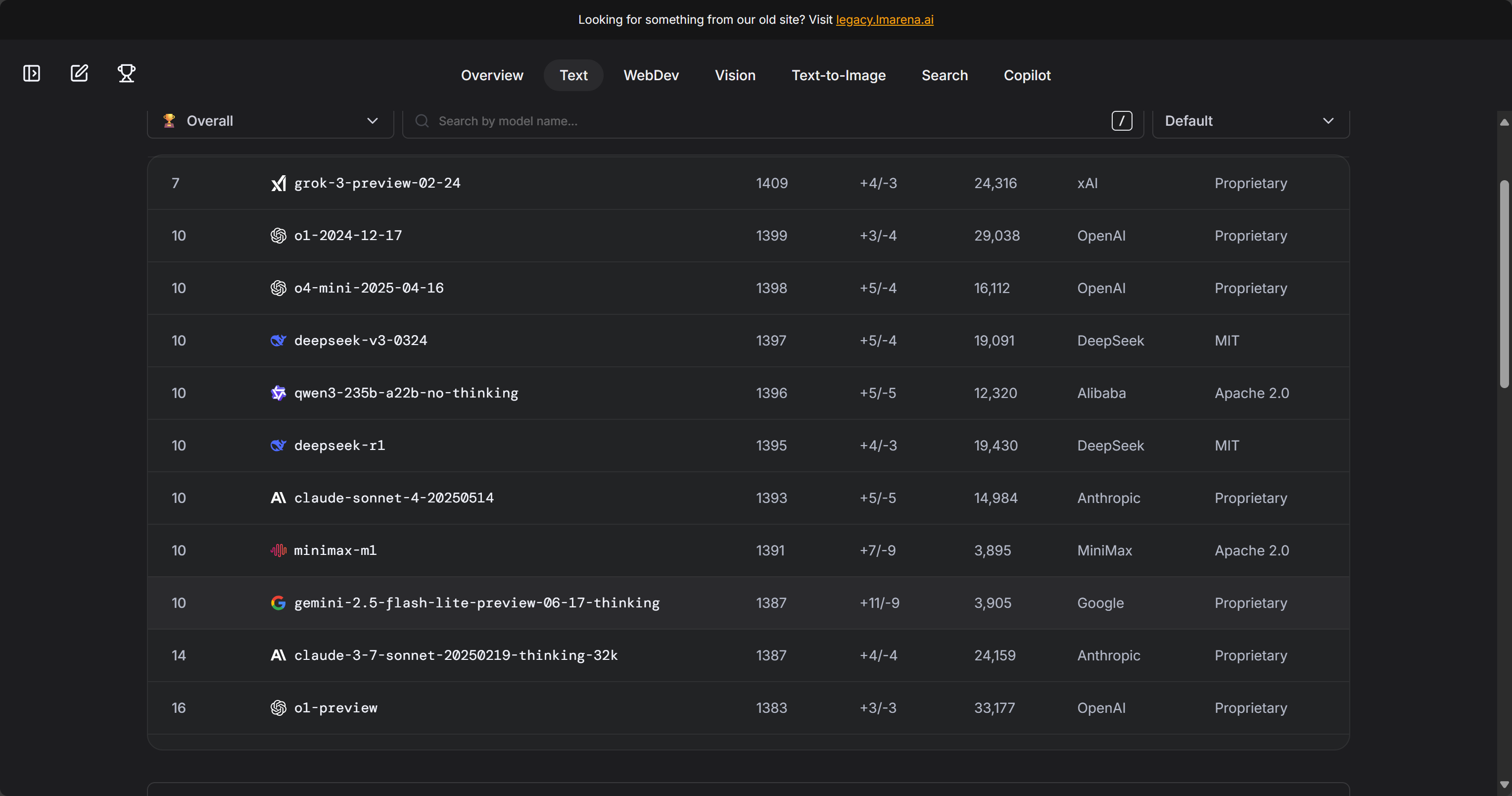
# Amazon

# Google





# OpenAI

# 建议

模型看新不看旧，看nova是去年12月出的，新模型即使参数量较小也能胜过一些参数量大的老模型（如前年出的GPT3.5，1750亿参数量），且Amazon大模型领域并不强（无论模型质量还是更新模型的频率），可以从上面排名看出，建议选择下面推荐的模型，价格差距不大，但能力远强于nova。

上面模型排名来自于最权威的模型匿名对战网站（[Overview Leaderboard | LMArena](https://lmarena.ai/leaderboard)），用户在不知道是哪个模型的情况下，选出哪个模型给的答案更好，具有较高的可信度。以Gemma-3-4B为参考基准，可以看出Nova-Pro的排名都低于Gemma-3-4B（Nova-Lite排名103，Nova-Pro排名85，Gemma-3-4B排名68）。

如果考虑成本，可以看下Google最新出的开源模型Gemma-3n-e4b（推荐，但需要自己部署到云服务器上），排名和实测能力超过Gemma-3-4B，参数量接近7B（磁盘占用4.24G），运行内存却接近Gemma-3-4B（做了优化）。（Gemma-3n-e4b排名68），部署成本应该比Nova-Lite更低，可以参考OpenRouter中的调用价格（Nova-Lite输入100M的Token要0.06美元，输出100M的Token要0.24美元，而Gemma-3n-e4b输入100M的Token要0.02美元，输出100M的Token要0.04美元）。

如果可以考虑略微比Nova稍微贵一点，但是能力顶尖的模型，有三款推荐：

* GPT-4.1-Flash-Lite输入100M的Token要0.1美元，输出100M的Token要0.4美元，排名47。
* Gemini-2.0-Flash-Lite输入100M的Token要0.075美元，输出100M的Token要0.3美元，排名30。（强烈推荐，只比nova贵一点点，能力很强，不需要自己部署）
* Gemini-2.5-Flash-Lite输入100M的Token要0.1美元，输出100M的Token要0.5美元，排名10。

最后说一句，强模型看似贵，但可能一次性就能解决问题，弱模型看似便宜，但可能要好几次旁敲侧击才能解决问题；同样难度问题，强模型用较少的token数量即可推出答案，弱模型要较多的token数量才有可能推出答案；此外强模型泛化能力要比弱模型强得多。价格差别不大的情况下，建议选强模型。（以上结论都进行了实测验证）

# 价格参考

[Models | OpenRouter](https://openrouter.ai/models?max_price=0.1)