EAGLE3 + Qwen3VL

训练：

路径：/mnt/beegfs/groups/cv/jiajun.lu/SpecForge_yuboyang

环境：/mnt/beegfs/groups/cv/yuboyang/anaconda3/envs/sglang

1. 数据预处理：

   /mnt/beegfs/groups/cv/jiajun.lu/SpecForge_yuboyang/scripts/fix_toon_data.py

2. 训练：./examples/run_weather.sh

3. 模型：./outputs/Qwen3-VL-nio-2b-eagle3/epoch_3

测试：

路径：/mnt/beegfs/groups/cv/jiajun.lu/sglang_yuboyang

环境：/mnt/beegfs/groups/cv/yuboyang/anaconda3/envs/sglang_eval

测试：./test_sglang_withbenchmark.py

实验结果汇总 EAGLE3

Base model = /mnt/beegfs/groups/cv/yuboyang/data/qwen3_nio_2b

测试集：_toon前100条数据

speculative-num-steps_speculative-eagle-topk_speculative-num-draft-tokens

Tree_structure	**prefill_latency** (s)	**spec_accept_rate** (%)	**e2e_latency** (s)	**compress ratio** (↑)
ep=9				
chain_structure 8node 8_1_8	0.21	31	0.64	3.47
chain_structure 5node	0.42	40	1.28	3.04
5_2_8	0.21	43	0.83	3.13

8_2_8	0.24	27.8	0.67	3.23
8_3_8	0.21	29.5	0.65	3.32
8_2_16	0.21	34.4	0.61	**3.73**
8_3_16	0.21	34.7	0.65	3.71
5_4_16	0.21	46.7	0.63	3.27
6_3_16	0.21	38.8	0.63	3.34
10_2_16	0.27	28	1.02	**3.86**
10_2_32	0.21	34	1.02	**4.31**
10_3_32	0.21	31	0.63	4.05
8_3_32	0.21	36	0.66	3.82
6_5_32	0.21	40	0.63	3.41
6_10_32	0.21	42	0.63	3.56
ep=3				
8_3_8				3.4
8_2_16				3.9
4_4_16（vs medusa）	0.21	51	0.63	3.02
4_2_14（vs medusa）	0.21	48	0.65	2.91
16_1_16				3.71
12_2_16	0.29	24.8	0.81	**3.93**
12_3_32				4.2
12_2_32				4.15
8_4_32				**4.17**
9_4_32				4.04

一个观察： 在加速比低的情况下，EAGLE3在sglang下跑出了210.39 tokens/s的decoding，远高于Medusa的43。证明对于2b级别的模型，推理框架对性能的提高超过了投机算法本身。

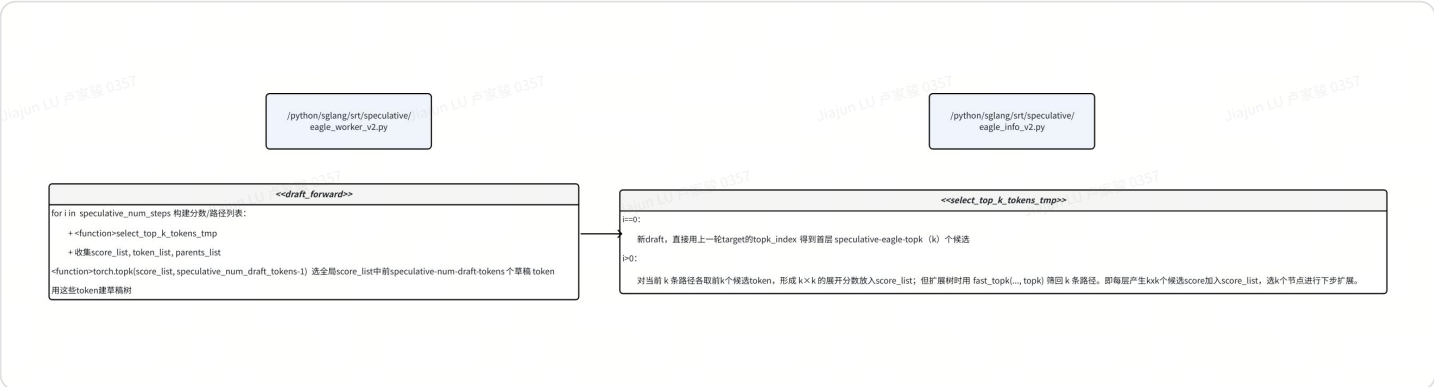
三个参数的具体作用

speculative-num-steps：可能的最大树深

speculative-eagle-topk：每层保留/向下扩展的路径数，展开时每条路径取 topk，形成 k×k 展开，再筛回 k 条路径继续。

speculative-num-draft-tokens：草稿树最终token数

具体流程：



对参数的理解

1. 简单数据，建议topk<=3，把更多节点分配给深度。

speculative-num-steps仅为理论最大树深，不代表实际。实际树深仅由全局选出speculative-num-draft-tokens个token形成的草稿树唯一决定。由于每个候选节点的 score = 根到该节点路径上各步的 softmax 概率之积，因此当topk过大时，会挤占token数，影响树深。

不挤占树深时，仍建议topk<=3。观察实验，加速比10_2_32 > 10_3_32。注意到这两种结构下node均盈余，树为每层2or3node、10层。而树每层多一个节点但加速比反降。WHY?

结论：topk != 树宽！ 最终树结构**由且仅由**全局score_list召回的前num-draft-tokens个token决定。注意到score_list中包含的是k + (num-steps-1)k*k个token。由概率的链乘特性，可能有：P（第i层的rank4节点）>P(第i+1层的rank3节点）。挤占名额，影响树深。