

Qwen Family Preliminary Benchmark Study

Revised statistical analysis report

This revised report removes the previous function-recommendations table because small task slices can overstate weak models, especially qwen3.5-0.8b. Function-level data is retained only as exploratory diagnostics.

Abstract

The experiment tested Qwen-family local models across model size, quantisation, KV-cache configuration, and MTP draft-token settings. The completed matrix covered 20 prompts across 120 model/configuration permutations for 2,400 prompt runs.

The clearest result is that model size drives most quality and pass-rate differences. qwen3.5-4b and qwen3.5-9b separate sharply from qwen3.5-0.8b and qwen3.5-2b, with qwen3.5-4b offering the most useful balance of quality, latency, and throughput for this workload.

MTP is not a universal quality knob. It changes latency and speed, and the best draft-token value depends on the model and configuration.

Testing Environment

Inference runtime: llama.cpp server build 9483 (5c394fdc8), built with Clang 19.1.5 for Windows x86_64.

GPU: NVIDIA GeForce RTX 4060 Ti with 16,380 MB VRAM. NVIDIA driver 591.74. CUDA 13.1.

Benchmark harness: Python 3.12.13. Configuration used an 8,192-token context, max_tokens 16,384, streaming enabled, thinking disabled, temperature 0.7, top_k 20, top_p 0.8, repeat_penalty 1.05, and a 300-second timeout.

Recommendations

- Use qwen3.5-4b Q8_0 mixed_q8k_q4nlv MTP=1 as the default balanced configuration: semantic 0.9125, pass 0.8500, elapsed 2.1977s.
- Use qwen3.5-4b Q8_0 mixed_q8k_q4nlv MTP=0 when quality is the priority and the extra latency is acceptable: semantic 0.9150, pass 0.8500, elapsed 2.7165s.
- Use qwen3.5-4b Q6_K q8 MTP=2 as the fastest acceptable configuration: semantic 0.8733, pass 0.8000, elapsed 1.5266s.
- Treat qwen3.5-9b as a specialist option rather than the default: its best run matched the top pass rate, but with higher average latency.
- Avoid qwen3.5-0.8b Q4_K_M mixed_q8k_q4nlv MTP=1 for this benchmark: semantic 0.6079, pass 0.3500.

- Do not use the function-level scores as deployment recommendations. They are exploratory diagnostics and need larger task samples before operational use.

Best Configurations

MODEL	QUANTIZATION	KV CACHE CONFIG	MTP DRAFT TOKENS	RUNS	AVG SEMANTIC SCORE	OVERALL PASS RATE	AVG ELAPSED S	AVG GENERATION TOK S	AVG PROMPT TOK S
qwen3.5-4b	Q8_0	mixed_q8k_q4nlv	1	20	0.9125	0.85	2.1977	60.55	1126.9
qwen3.5-9b	Q8_0	q8	4	20	0.9125	0.85	2.7197	51.61	1008.8
qwen3.5-4b	Q8_0	mixed_q8k_q4nlv	0	20	0.915	0.85	2.7165	43.70	1183.4
qwen3.5-4b	Q4_K_M	q8	1	20	0.8967	0.8	1.5285	88.39	1668.1
qwen3.5-4b	Q6_K	q8	1	20	0.8942	0.8	1.8037	78.12	1549.6
qwen3.5-4b	Q8_0	q8	2	20	0.8917	0.8	1.6557	76.11	1481.0
qwen3.5-4b	Q6_K	q8	3	20	0.8825	0.8	1.6206	87.24	1524.5
qwen3.5-9b	Q8_0	mixed_q8k_q4nlv	1	20	0.9025	0.85	3.1701	38.49	845.6
qwen3.5-4b	Q4_K_M	mixed_q8k_q4nlv	0	20	0.9042	0.8	2.395	62.31	1306.3
qwen3.5-9b	Q6_K	q8	2	20	0.9025	0.8	2.2967	59.51	1078.7
qwen3.5-9b	Q4_K_M	mixed_q8k_q4nlv	0	20	0.9008	0.85	3.8691	39.83	1024.0
qwen3.5-9b	Q6_K	q8	0	20	0.9021	0.85	3.5504	34.58	1198.1
qwen3.5-4b	Q6_K	q8	2	20	0.8733	0.8	1.5266	87.44	1514.0
qwen3.5-9b	Q8_0	mixed_q8k_q4nlv	2	20	0.89	0.85	2.8886	46.13	834.3
qwen3.5-4b	Q8_0	q8	1	20	0.8917	0.8	2.0461	66.07	1515.1
qwen3.5-4b	Q4_K_M	q8	3	20	0.8842	0.75	1.4458	96.69	1653.7
qwen3.5-4b	Q4_K_M	mixed_q8k_q4nlv	2	20	0.8967	0.75	1.9135	86.40	1212.0
qwen3.5-9b	Q4_K_M	q8	2	20	0.8858	0.8	1.9622	68.57	1199.3
qwen3.5-9b	Q4_K_M	q8	4	20	0.8942	0.8	3.3488	71.28	1199.7
qwen3.5-9b	Q6_K	mixed_q8k_q4nlv	3	20	0.8942	0.8	2.3105	57.72	868.8

20 of 120 rows shown.

Main Effects

Model main effect

MODEL	RUNS	AVG SEMANTIC SCORE	OVERALL PASS RATE	AVG ELAPSED S	AVG GENERATION TOK S	AVG PROMPT TOK S
qwen3.5-0.8b	600	0.7309	0.5117	0.9659	144.8	3434.1
qwen3.5-2b	600	0.772	0.615	1.3081	108.4	2738.8
qwen3.5-4b	600	0.8775	0.775	1.8913	74.36	1384.8
qwen3.5-9b	600	0.8818	0.7767	3.7623	51.41	1012.5

Quantisation main effect

QUANTIZATION	RUNS	AVG SEMANTIC SCORE	OVERALL PASS RATE	AVG ELAPSED S	AVG GENERATION TOK S	AVG PROMPT TOK S	AVG DRAFT ACCEPT PCT	AVG DRAFT REJECT PCT	AVG SEMANTIC POINTS PER S
Q4_K_M	800	0.7985	0.6388	1.7881	103.2	2211.7	0.5623	0.4651	4.4061
Q6_K	800	0.8239	0.6913	1.9415	96.64	2123.6	0.5766	0.4418	4.3786
Q8_0	800	0.8243	0.6787	2.216	84.53	2094.7	0.583	0.4436	3.9542

KV-cache main effect

KV CACHE CONFIG	RUNS	AVG SEMANTIC SCORE	OVERALL PASS RATE	AVG ELAPSED S	AVG GENERATION TOK S	AVG PROMPT TOK S	AVG DRAFT ACCEPT PCT	AVG DRAFT REJECT PCT	AVG SEMANTIC POINTS PER S
mixed_q8k_q4nlv	1200	0.8163	0.6675	2.0858	88.54	1802.5	0.5752	0.4473	3.997
q8	1200	0.8148	0.6717	1.878	101.0	2483.6	0.5727	0.453	4.4956

MTP draft-token main effect

MTP DRAFT TOKENS	RUNS	AVG SEMANTIC SCORE	OVERALL PASS RATE	AVG ELAPSED S	AVG GENERATION TOK S	AVG PROMPT TOK S	AVG DRAFT ACCEPT PCT	AVG DRAFT REJECT PCT	AVG SEMANTIC POINTS PER S
0	480	0.8192	0.675	2.6227	95.54	2374.2			4.2019
1	480	0.8162	0.6667	1.9261	101.4	2119.6	0.7592	0.2545	4.4552
2	480	0.8126	0.6687	1.6865	99.78	2090.0	0.6134	0.4009	4.3495
3	480	0.8152	0.6646	1.638	92.85	2085.3	0.5053	0.5249	4.2546

MTP DRAFT TOKENS	RUNS	AVG SEMANTIC SCORE	OVERALL PASS RATE	AVG ELAPSED S	AVG GENERATION TOK S	AVG PROMPT TOK S	AVG DRAFT ACCEPT PCT	AVG DRAFT REJECT PCT	AVG SEMANTIC POINTS PER S
4	480	0.8145	0.6729	2.036	84.27	2048.6	0.418	0.6204	3.9702

Function-area main effect

FUNCTION AREA	RUNS	AVG SEMANTIC SCORE	OVERALL PASS RATE	AVG ELAPSED S	AVG GENERATION TOK S	AVG PROMPT TOK S	AVG DRAFT ACCEPT PCT	AVG DRAFT REJECT PCT	AVG SEMANTIC POINTS PER S
Customer Support	120	0.6333	0.2167	1.4037	94.72	2060.9	0.5403	0.4444	3.38
Data/Analytics	240	0.7935	0.5042	2.7715	90.67	2079.8	0.5068	0.495	3.3198
Finance	360	0.7069	0.5389	3.6445	95.21	2660.5	0.5944	0.3628	3.7239
HR	240	0.8917	0.8	0.2368	98.36	1917.7	0.6707	0.4423	8.8507
IT	240	0.8719	0.8708	1.1954	96.61	1955.5	0.5666	0.4411	5.4666
Knowledge Management	240	0.9156	0.7542	1.9914	89.56	1870.4	0.4823	0.4708	2.7336
Legal	240	0.8108	0.725	2.3235	93.32	2058.2	0.5566	0.452	3.287
Operations	480	0.77	0.6146	1.7889	94.54	2042.5	0.5688	0.4822	3.5426
Sales	240	0.9549	0.8958	1.5538	99.98	2450.0	0.6574	0.4697	4.4443

ANOVA And Effect Sizes

Semantic score ANOVA

EFFECT	F	P VALUE	ETA SQUARED	PARTIAL ETA SQUARED
C(model)	52.86	0	0.0624	0.0628
C(quantization)	2.6788	0.0689	0.0021	0.0023
C(kv_cache_config)	0.0226	0.8806	0	0
C(mtp_draft_tokens_factor)	0.0438	0.9964	0.0001	0.0001
C(model):C(mtp_draft_tokens_factor)	0.2968	0.99	0.0014	0.0015
C(model):C(quantization)	0.9042	0.4908	0.0021	0.0023
C(kv_cache_config):C(mtp_draft_tokens_factor)	0.1925	0.9424	0.0003	0.0003
Residual			0.9316	

Elapsed seconds ANOVA

EFFECT	F	P VALUE	ETA SQUARED	PARTIAL ETA SQUARED
C(model)	48.58	0	0.0565	0.058
C(quantization)	1.9579	0.1414	0.0015	0.0017
C(kv_cache_config)	1.3488	0.2456	0.0005	0.0006
C(mtp_draft_tokens_factor)	3.8883	0.0038	0.006	0.0065
C(model):C(mtp_draft_tokens_factor)	3.4208	0.0001	0.0159	0.017
C(model):C(quantization)	0.8223	0.5525	0.0019	0.0021
C(kv_cache_config):C(mtp_draft_tokens_factor)	0.4781	0.7519	0.0007	0.0008
Residual			0.9169	

Generation tokens/sec ANOVA

EFFECT	F	P VALUE	ETA SQUARED	PARTIAL ETA SQUARED
C(model)	7508.1	0	0.721	0.905
C(quantization)	537.0	0	0.0344	0.3123
C(kv_cache_config)	710.3	0	0.0227	0.231
C(mtp_draft_tokens_factor)	166.2	0	0.0213	0.2194
C(model):C(mtp_draft_tokens_factor)	321.6	0	0.1235	0.62
C(model):C(quantization)	5.5278	0	0.0011	0.0138
C(kv_cache_config):C(mtp_draft_tokens_factor)	2.8373	0.0231	0.0004	0.0048
Residual			0.0757	

Overall pass ANOVA

EFFECT	F	P VALUE	ETA SQUARED	PARTIAL ETA SQUARED
C(model)	48.04	0	0.0571	0.0574
C(quantization)	2.8622	0.0573	0.0023	0.0024
C(kv_cache_config)	0.0496	0.8239	0	0
C(mtp_draft_tokens_factor)	0.0426	0.9966	0.0001	0.0001
C(model):C(mtp_draft_tokens_factor)	0.2197	0.9976	0.001	0.0011
C(model):C(quantization)	0.8986	0.4949	0.0021	0.0023
C(kv_cache_config):C(mtp_draft_tokens_factor)	0.2032	0.9367	0.0003	0.0003
Residual			0.9371	

MTP And Interaction Effects

Global MTP sweet spot

MTP DRAFT TOKENS	RUNS	AVG SEMANTIC SCORE	OVERALL PASS RATE	AVG ELAPSED S	AVG GENERATION TOK S	AVG PROMPT TOK S	AVG DRAFT ACCEPT PCT	AVG DRAFT REJECT PCT	AVG SEMANTIC POINTS PER S	SELECTION RULE
1	480	0.8162	0.6667	1.9261	101.4	2119.6	0.7592	0.2545	4.4552	within 2pp pass and 0.02 semantic, then semantic points/s, rejection, lower draft

MTP sweet spot by model

MODEL	MTP DRAFT TOKENS	RUNS	AVG SEMANTIC SCORE	OVERALL PASS RATE	AVG ELAPSED S	AVG GENERATION TOK S	AVG PROMPT TOK S	AVG DRAFT ACCEPT PCT	AVG DRAFT REJECT PCT	AVG SEMANTIC POINTS PER S	SELECTION RULE
qwen3.5-0.8b	3	120	0.7563	0.5417	1.0986	125.0	3345.6	0.2186	0.7837	5.6533	within 2pp pass and 0.02 semantic, then semantic points/s, rejection, lower draft
qwen3.5-2b	2	120	0.7781	0.6333	1.2151	117.1	2663.2	0.5831	0.4466	4.9031	within 2pp pass and 0.02 semantic, then semantic points/s, rejection, lower draft
qwen3.5-4b	3	120	0.8702	0.775	1.7952	82.42	1352.6	0.6493	0.4077	3.9874	within 2pp pass and 0.02 semantic, then semantic points/s, rejection, lower draft
qwen3.5-9b	4	120	0.8825	0.7917	3.7843	60.05	982.3	0.6512	0.4496	2.9258	within 2pp pass and 0.02 semantic, then semantic points/s, rejection, lower draft

MTP sweet spot by quantisation

QUANTIZATION	MTP DRAFT TOKENS	RUNS	AVG SEMANTIC SCORE	OVERALL PASS RATE	AVG ELAPSED S	AVG GENERATION TOK S	AVG PROMPT TOK S	AVG DRAFT ACCEPT PCT	AVG DRAFT REJECT PCT	AVG SEMANTIC POINTS PER S	SELECTION RULE
Q4_K_M	0	160	0.8208	0.6813	2.1841	107.4	2428.3			4.5482	within 2pp pass and 0.02 semantic, then semantic points/s, rejection, lower draft
Q6_K	2	160	0.8338	0.7125	1.4938	102.6	2068.8	0.6244	0.3848	4.6952	within 2pp pass and 0.02 semantic, then semantic points/s, rejection, lower draft
Q8_0	1	160	0.8306	0.7	2.0884	90.51	2070.4	0.7667	0.2518	4.1275	within 2pp pass and 0.02 semantic, then semantic points/s, rejection, lower draft

MTP sweet spot by KV-cache

KV CACHE CONFIG	MTP DRAFT TOKENS	RUNS	AVG SEMANTIC SCORE	OVERALL PASS RATE	AVG ELAPSED S	AVG GENERATION TOK S	AVG PROMPT TOK S	AVG DRAFT ACCEPT PCT	AVG DRAFT REJECT PCT	AVG SEMANTIC POINTS PER S	SELECTION RULE
mixed_q8k_q4nlv	3	240	0.816	0.6708	1.7118	87.10	1760.2	0.5079	0.5241	3.998	within 2pp pass and 0.02 semantic, then semantic points/s, rejection, lower draft
q8	1	240	0.8117	0.6708	1.8576	108.4	2446.8	0.761	0.2567	4.7209	within 2pp pass and 0.02 semantic, then semantic points/s, rejection, lower draft

Model by draft-token interaction

MODEL	MTP DRAFT TOKENS	RUNS	AVG SEMANTIC SCORE	OVERALL PASS RATE	AVG ELAPSED S	AVG GENERATION TOK S	AVG PROMPT TOK S	AVG DRAFT ACCEPT PCT	AVG DRAFT REJECT PCT	AVG SEMANTIC POINTS PER S
qwen3.5-0.8b	0	120	0.7419	0.5167	0.7636	182.3	3808.5			7.2598
qwen3.5-0.8b	1	120	0.7134	0.5	0.809	165.8	3404.8	0.557	0.4174	6.6824
qwen3.5-0.8b	2	120	0.7153	0.4917	0.9511	143.1	3354.3	0.3122	0.6748	5.8235
qwen3.5-0.8b	3	120	0.7563	0.5417	1.0986	125.0	3345.6	0.2186	0.7837	5.6533

MODEL	MTP DRAFT TOKENS	RUNS	AVG SEMANTIC SCORE	OVERALL PASS RATE	AVG ELAPSED S	AVG GENERATION TOK S	AVG PROMPT TOK S	AVG DRAFT ACCEPT PCT	AVG DRAFT REJECT PCT	AVG SEMANTIC POINTS PER S
qwen3.5-0.8b	4	120	0.7275	0.5083	1.207	108.0	3257.4	0.1598	0.8352	4.9396
qwen3.5-2b	0	120	0.7643	0.6083	1.2039	109.0	3059.8			4.8306
qwen3.5-2b	1	120	0.7793	0.6083	1.4203	118.8	2702.6	0.7444	0.2987	5.1551
qwen3.5-2b	2	120	0.7781	0.6333	1.2151	117.1	2663.2	0.5831	0.4466	4.9031
qwen3.5-2b	3	120	0.7603	0.6083	1.3079	105.0	2651.6	0.4243	0.6102	4.4808
qwen3.5-2b	4	120	0.7779	0.6167	1.3932	92.01	2616.8	0.3249	0.6995	4.2282
qwen3.5-4b	0	120	0.8835	0.7833	2.3499	56.11	1517.5			2.9099
qwen3.5-4b	1	120	0.8894	0.7833	1.8556	73.74	1374.2	0.8674	0.139	3.588
qwen3.5-4b	2	120	0.8743	0.7583	1.6964	82.38	1343.0	0.758	0.2512	3.907
qwen3.5-4b	3	120	0.8702	0.775	1.7952	82.42	1352.6	0.6493	0.4077	3.9874
qwen3.5-4b	4	120	0.8702	0.775	1.7595	77.00	1338.0	0.5361	0.4973	3.7872
qwen3.5-9b	0	120	0.8873	0.7917	6.1736	33.94	1093.1			1.8074
qwen3.5-9b	1	120	0.8828	0.775	3.6196	47.47	996.8	0.868	0.1629	2.3954
qwen3.5-9b	2	120	0.8826	0.7917	2.8834	56.51	999.4	0.8004	0.2309	2.7644
qwen3.5-9b	3	120	0.874	0.7333	2.3503	58.95	991.4	0.7289	0.298	2.8968
qwen3.5-9b	4	120	0.8825	0.7917	3.7843	60.05	982.3	0.6512	0.4496	2.9258

Quantisation by draft-token interaction

QUANTIZATION	MTP DRAFT TOKENS	RUNS	AVG SEMANTIC SCORE	OVERALL PASS RATE	AVG ELAPSED S	AVG GENERATION TOK S	AVG PROMPT TOK S	AVG DRAFT ACCEPT PCT	AVG DRAFT REJECT PCT	AVG SEMANTIC POINTS PER S
Q4_K_M	0	160	0.8208	0.6813	2.1841	107.4	2428.3			4.5482
Q4_K_M	1	160	0.7973	0.6375	1.6576	110.9	2188.0	0.7493	0.2687	4.6519
Q4_K_M	2	160	0.7772	0.625	1.4888	107.8	2158.3	0.5972	0.4151	4.3051
Q4_K_M	3	160	0.7952	0.6188	1.4675	99.05	2161.4	0.4923	0.5314	4.365
Q4_K_M	4	160	0.8017	0.6312	2.1424	90.61	2122.6	0.4104	0.6454	4.1604
Q6_K	0	160	0.8131	0.6813	2.568	96.79	2379.6			4.2086

QUANTIZATION	MTP DRAFT TOKENS	RUNS	AVG SEMANTIC SCORE	OVERALL PASS RATE	AVG ELAPSED S	AVG GENERATION TOK S	AVG PROMPT TOK S	AVG DRAFT ACCEPT PCT	AVG DRAFT REJECT PCT	AVG SEMANTIC POINTS PER S
Q6_K	1	160	0.8206	0.6625	2.0323	102.9	2100.3	0.7616	0.243	4.5863
Q6_K	2	160	0.8338	0.7125	1.4938	102.6	2068.8	0.6244	0.3848	4.6952
Q6_K	3	160	0.8226	0.6937	1.5438	95.40	2060.7	0.5033	0.5166	4.3383
Q6_K	4	160	0.8294	0.7063	2.0698	85.52	2011.7	0.417	0.6228	4.0644
Q8_0	0	160	0.8237	0.6625	3.1161	82.40	2314.7			3.849
Q8_0	1	160	0.8306	0.7	2.0884	90.51	2070.4	0.7667	0.2518	4.1275
Q8_0	2	160	0.8268	0.6687	2.077	88.96	2042.8	0.6186	0.4027	4.0482
Q8_0	3	160	0.8279	0.6813	1.9027	84.10	2033.8	0.5203	0.5267	4.0604
Q8_0	4	160	0.8124	0.6813	1.8959	76.68	2011.6	0.4265	0.593	3.6859

KV-cache by draft-token interaction

KV CACHE CONFIG	MTP DRAFT TOKENS	RUNS	AVG SEMANTIC SCORE	OVERALL PASS RATE	AVG ELAPSED S	AVG GENERATION TOK S	AVG PROMPT TOK S	AVG DRAFT ACCEPT PCT	AVG DRAFT REJECT PCT	AVG SEMANTIC POINTS PER S
mixed_q8k_q4nlv	0	240	0.8231	0.6833	2.966	88.34	1947.3			3.9142
mixed_q8k_q4nlv	1	240	0.8207	0.6625	1.9946	94.53	1792.3	0.7574	0.2524	4.1896
mixed_q8k_q4nlv	2	240	0.8153	0.6625	1.6724	93.62	1778.0	0.6175	0.398	4.1903
mixed_q8k_q4nlv	3	240	0.816	0.6708	1.7118	87.10	1760.2	0.5079	0.5241	3.998
mixed_q8k_q4nlv	4	240	0.8065	0.6583	2.084	79.11	1736.0	0.4181	0.6147	3.6929
q8	0	240	0.8153	0.6667	2.2795	102.7	2797.5			4.4896
q8	1	240	0.8117	0.6708	1.8576	108.4	2446.8	0.761	0.2567	4.7209
q8	2	240	0.8098	0.675	1.7007	105.9	2401.9	0.6094	0.4037	4.5087
q8	3	240	0.8145	0.6583	1.5642	98.60	2410.4	0.5026	0.5257	4.5112
q8	4	240	0.8225	0.6875	1.988	89.43	2361.3	0.4179	0.6261	4.2475

Model by quantisation interaction

MODEL	QUANTIZATION	RUNS	AVG SEMANTIC SCORE	OVERALL PASS RATE	AVG ELAPSED S	AVG GENERATION TOK S	AVG PROMPT TOK S	AVG DRAFT ACCEPT PCT	AVG DRAFT REJECT PCT	AVG SEMANTIC POINTS PER S
qwen3.5-0.8b	Q4_K_M	200	0.6898	0.46	0.9853	151.2	3458.0	0.2918	0.6965	5.8313
qwen3.5-0.8b	Q6_K	200	0.7629	0.585	0.9264	148.7	3422.9	0.3183	0.667	6.4542
qwen3.5-0.8b	Q8_0	200	0.7398	0.49	0.9859	134.6	3421.5	0.3256	0.6698	5.9296
qwen3.5-2b	Q4_K_M	200	0.7539	0.585	1.3261	118.4	2822.8	0.5116	0.5317	4.864
qwen3.5-2b	Q6_K	200	0.7749	0.625	1.2094	110.6	2712.7	0.5222	0.5019	4.8281
qwen3.5-2b	Q8_0	200	0.7871	0.635	1.3888	96.08	2680.9	0.5237	0.5076	4.4666
qwen3.5-4b	Q4_K_M	200	0.8769	0.755	1.7047	83.16	1464.2	0.6882	0.3376	4.0289
qwen3.5-4b	Q6_K	200	0.8725	0.77	1.8091	75.04	1361.2	0.7044	0.3146	3.6344
qwen3.5-4b	Q8_0	200	0.8831	0.8	2.1601	64.89	1329.0	0.7156	0.3193	3.2444
qwen3.5-9b	Q4_K_M	200	0.8732	0.755	3.1362	59.85	1101.8	0.7577	0.2948	2.9003
qwen3.5-9b	Q6_K	200	0.8853	0.785	3.8213	51.84	988.2	0.7615	0.2837	2.5975
qwen3.5-9b	Q8_0	200	0.8871	0.79	4.3293	42.56	947.3	0.7671	0.2776	2.1761

Model by KV-cache interaction

MODEL	KV CACHE CONFIG	RUNS	AVG SEMANTIC SCORE	OVERALL PASS RATE	AVG ELAPSED S	AVG GENERATION TOK S	AVG PROMPT TOK S	AVG DRAFT ACCEPT PCT	AVG DRAFT REJECT PCT	AVG SEMANTIC POINTS PER S
qwen3.5-0.8b	mixed_q8k_q4nlv	300	0.7333	0.5133	1.0546	131.6	2789.7	0.3099	0.6842	5.6087
qwen3.5-0.8b	q8	300	0.7284	0.51	0.8771	158.1	4078.6	0.314	0.6713	6.5347
qwen3.5-2b	mixed_q8k_q4nlv	300	0.7766	0.62	1.3423	102.4	2353.9	0.5233	0.5094	4.4877
qwen3.5-2b	q8	300	0.7674	0.61	1.2739	114.3	3123.8	0.5151	0.518	4.9515
qwen3.5-4b	mixed_q8k_q4nlv	300	0.8779	0.7733	1.9881	70.26	1166.8	0.7008	0.3222	3.438
qwen3.5-4b	q8	300	0.8771	0.7767	1.7945	78.45	1602.1	0.7046	0.3254	3.8339
qwen3.5-9b	mixed_q8k_q4nlv	300	0.8775	0.7633	3.958	49.70	894.6	0.7669	0.2732	2.4536
qwen3.5-9b	q8	300	0.8862	0.79	3.5665	53.13	1129.9	0.7573	0.2975	2.6624

Enterprise Function Suitability

The function recommendation table has been intentionally removed. The matrix below is retained only as an exploratory diagnostic and should not be read as a recommendation for qwen3.5-0.8b or any other model.

MODEL	QUANTIZATION	KV CACHE CONFIG	MTP DRAFT TOKENS	FINANCE	HR	IT	CUSTOMER SUPPORT	LEGAL	SALES	OPERATIONS	DATA ANALYTICS	KNOWLEDGE MANAGEMENT	OVERALL
qwen3.5-0.8b	Q4_K_M	mixed_q8k_q4nlv	0	0.8333	0.75	0.9167	0.1667	0.8	0.9167	0.7708	0.875	1	0.8133
qwen3.5-0.8b	Q4_K_M	mixed_q8k_q4nlv	1	0.75	0.5	0.75	0.5	0.2	0.8333	0.5729	0.65	0.625	0.6079
qwen3.5-0.8b	Q4_K_M	mixed_q8k_q4nlv	2	0.5	0.5	0.9167	0.5	0.4	0.6667	0.5764	0.875	0.875	0.6386
qwen3.5-0.8b	Q4_K_M	mixed_q8k_q4nlv	3	0.5	0.75	0.9167	0.1667	0.6	0.9167	0.6354	0.75	0.875	0.6912
qwen3.5-0.8b	Q4_K_M	mixed_q8k_q4nlv	4	0.5	0.5	0.8333	0.6667	0.8	0.9167	0.5208	0.525	0.75	0.645
qwen3.5-0.8b	Q4_K_M	q8	0	0.8333	0.75	0.4333	0.6667	0.2	0.8333	0.8021	0.775	0.75	0.6929
qwen3.5-0.8b	Q4_K_M	q8	1	0.5	0.75	0.9167	0.5	0.9	0.9167	0.5347	0.625	0.875	0.7053
qwen3.5-0.8b	Q4_K_M	q8	2	0.0833	0.75	0.8333	0.6667	0.7	0.8333	0.4896	0.625	1	0.6179
qwen3.5-0.8b	Q4_K_M	q8	3	0.8333	0.75	0.7333	0.6667	0.7	1	0.6042	0.75	0.875	0.76
qwen3.5-0.8b	Q4_K_M	q8	4	0.75	0.75	0.8167	0.5	0.4	0.9167	0.6875	0.75	0.875	0.7258
qwen3.5-0.8b	Q6_K	mixed_q8k_q4nlv	0	0.4167	0.75	0.9167	0.5	0.8	0.9167	0.7118	0.625	0.625	0.6932
qwen3.5-0.8b	Q6_K	mixed_q8k_q4nlv	1	0.8333	0.75	0.8333	0.5	0.8	0.9167	0.4792	0.75	0.625	0.7133
qwen3.5-0.8b	Q6_K	mixed_q8k_q4nlv	2	0.8333	0.75	0.9167	0.6667	0.6	0.9167	0.7708	0.65	0.875	0.7833
qwen3.5-0.8b	Q6_K	mixed_q8k_q4nlv	3	0.5	1	0.9167	1	0.6	0.9167	0.691	0.875	0.75	0.769
qwen3.5-0.8b	Q6_K	mixed_q8k_q4nlv	4	0.75	0.75	0.9167	0.6667	0.9	1	0.8056	0.875	0.875	0.8386
qwen3.5-0.8b	Q6_K	q8	0	0.75	0.75	0.9167	0.6667	0.7	1	0.6562	0.65	0.75	0.7538
qwen3.5-0.8b	Q6_K	q8	1	0.75	0.75	0.9167	0.6667	0.9	0.9167	0.6667	0.65	0.75	0.7675
qwen3.5-0.8b	Q6_K	q8	2	0.8333	0.75	1	0.8333	0.4	0.9167	0.6875	1	0.75	0.7858
qwen3.5-0.8b	Q6_K	q8	3	0.75	0.25	0.9167	0.6667	1	1	0.7431	0.75	1	0.7861
qwen3.5-0.8b	Q6_K	q8	4	0.75	0.75	0.9167	0.6667	0.3	0.9167	0.7083	0.875	0.75	0.7383
qwen3.5-0.8b	Q8_0	mixed_q8k_q4nlv	0	0.9167	0.75	0.9167	0.5	0.7	0.6667	0.6875	0.75	0.875	0.7658
qwen3.5-0.8b	Q8_0	mixed_q8k_q4nlv	1	0.9167	0.75	0.9167	0.5	0.8	0.75	0.8333	0.75	0.875	0.8133
qwen3.5-0.8b	Q8_0	mixed_q8k_q4nlv	2	0.75	0.75	0.8333	0.5	0.7	1	0.7708	0.75	0.75	0.77

MODEL	QUANTIZATION	KV CACHE CONFIG	MTP DRAFT TOKENS	FINANCE	HR	IT	CUSTOMER SUPPORT	LEGAL	SALES	OPERATIONS	DATA ANALYTICS	KNOWLEDGE MANAGEMENT	OVERALL
qwen3.5-0.8b	Q8_0	mixed_q8k_q4nlv	3	0.75	0.75	0.9167	0.6667	0.7	0.9167	0.6875	0.875	1	0.7992
qwen3.5-0.8b	Q8_0	mixed_q8k_q4nlv	4	0.75	0.75	0.75	0.6667	0.2	0.9167	0.7743	0.325	0.625	0.6574
qwen3.5-0.8b	Q8_0	q8	0	0.75	0.75	0.9167	0.8333	0.3	0.9167	0.5729	0.875	0.875	0.7321
qwen3.5-0.8b	Q8_0	q8	1	0.4167	0.75	0.9167	0.5	0.2	0.9167	0.6597	0.875	0.875	0.6728
qwen3.5-0.8b	Q8_0	q8	2	0.75	0.75	0.8333	0.8333	0.5	0.9167	0.5208	0.75	0.625	0.6958
qwen3.5-0.8b	Q8_0	q8	3	0.75	0.75	0.8333	0.6667	0.7	1	0.6042	0.875	0.5	0.7325
qwen3.5-0.8b	Q8_0	q8	4	0.8333	0.75	0.9167	0.5	0.7	0.9167	0.7188	0.75	0.625	0.7596
qwen3.5-2b	Q4_K_M	mixed_q8k_q4nlv	0	0.5	0.75	1	0.6667	0.7	0.6667	0.7743	0.875	0.875	0.7499
qwen3.5-2b	Q4_K_M	mixed_q8k_q4nlv	1	0.5	1	1	0.5	0.7	0.9167	0.8333	0.75	0.875	0.7908
qwen3.5-2b	Q4_K_M	mixed_q8k_q4nlv	2	0.5	1	1	0.6667	0.9	0.9167	0.6285	0.875	0.5	0.7532
qwen3.5-2b	Q4_K_M	mixed_q8k_q4nlv	3	0.25	0.75	0.9	0.6667	0.9	1	0.8333	0.875	0.625	0.7425
qwen3.5-2b	Q4_K_M	mixed_q8k_q4nlv	4	0.5	0.75	0.9167	0.6667	0.6	0.9167	0.8333	0.75	0.875	0.7558
qwen3.5-2b	Q4_K_M	q8	0	0.5	0.75	0.7167	0.1667	0.9	1	0.7431	0.875	1	0.7561
qwen3.5-2b	Q4_K_M	q8	1	0.5	0.75	0.9167	0.6667	0.7	0.9167	0.8333	0.875	0.75	0.7658
qwen3.5-2b	Q4_K_M	q8	2	0.5	1	0.8167	0.5	0.6	0.9167	0.7743	0.75	0.75	0.7382
qwen3.5-2b	Q4_K_M	q8	3	0.1667	0.75	0.8167	0.6667	0.9	0.9167	0.8021	0.75	0.875	0.7196
qwen3.5-2b	Q4_K_M	q8	4	0.5833	0.75	0.7333	0.6667	0.7	0.9167	0.7431	0.875	1	0.7669
qwen3.5-2b	Q6_K	mixed_q8k_q4nlv	0	0.5	0.75	0.9	0.5	0.9	0.9167	0.8056	0.75	1	0.7828
qwen3.5-2b	Q6_K	mixed_q8k_q4nlv	1	0.5	1	0.9167	0.6667	0.9	0.9167	0.8333	0.875	0.75	0.8108
qwen3.5-2b	Q6_K	mixed_q8k_q4nlv	2	0.1667	1	0.9	0.6667	1	0.9167	0.8333	0.75	1	0.7817
qwen3.5-2b	Q6_K	mixed_q8k_q4nlv	3	0.5	0.75	0.9167	0.6667	0.8	0.9167	0.8333	0.875	1	0.8008
qwen3.5-2b	Q6_K	mixed_q8k_q4nlv	4	0.5	0.75	0.8167	0.5	0.8	1	0.7708	0.875	1	0.7783
qwen3.5-2b	Q6_K	q8	0	0.25	0.75	1	0.6667	0.7	0.9167	0.8333	0.75	0.875	0.7367
qwen3.5-2b	Q6_K	q8	1	0.5	1	0.7167	0.5	0.8	0.9167	0.6667	0.875	0.875	0.7517
qwen3.5-2b	Q6_K	q8	2	0.5833	1	0.9167	0.5	0.9	0.8333	0.8333	0.625	1	0.8067
qwen3.5-2b	Q6_K	q8	3	0.1667	0.75	1	0.5	0.8	0.9167	0.8021	0.875	0.75	0.7196
qwen3.5-2b	Q6_K	q8	4	0.5833	0.75	1	0.1667	0.7	0.9167	0.8021	0.875	1	0.7804

MODEL	QUANTIZATION	KV CACHE CONFIG	MTP DRAFT TOKENS	FINANCE	HR	IT	CUSTOMER SUPPORT	LEGAL	SALES	OPERATIONS	DATA ANALYTICS	KNOWLEDGE MANAGEMENT	OVERALL
qwen3.5-2b	Q8_0	mixed_q8k_q4nlv	0	0.25	0.75	0.9167	0.5	0.8	0.9167	0.8056	0.75	1	0.7369
qwen3.5-2b	Q8_0	mixed_q8k_q4nlv	1	0.5	1	0.8333	0.5	0.8	1	0.8056	0.75	0.875	0.7869
qwen3.5-2b	Q8_0	mixed_q8k_q4nlv	2	0.5833	1	0.9	0.6667	0.8	1	0.8333	0.75	0.875	0.82
qwen3.5-2b	Q8_0	mixed_q8k_q4nlv	3	0.5	1	1	0.5	0.9	0.8333	0.8056	0.625	0.75	0.7719
qwen3.5-2b	Q8_0	mixed_q8k_q4nlv	4	0.5	0.75	0.9	0.5	0.8	1	0.8333	0.875	0.875	0.7867
qwen3.5-2b	Q8_0	q8	0	0.5	1	1	0.6667	0.9	0.8333	0.8333	0.75	1	0.8233
qwen3.5-2b	Q8_0	q8	1	0.5	0.75	0.8167	0.6667	0.9	0.9167	0.8021	0.75	0.875	0.7696
qwen3.5-2b	Q8_0	q8	2	0.25	0.75	1	0.6667	0.9	0.9167	0.8333	0.875	0.875	0.7692
qwen3.5-2b	Q8_0	q8	3	0.5	1	0.8	0.3333	0.9	0.9167	0.8333	0.875	1	0.8075
qwen3.5-2b	Q8_0	q8	4	0.5833	0.75	0.9	0.6667	0.8	0.9167	0.8333	0.75	1	0.7992
qwen3.5-4b	Q4_K_M	mixed_q8k_q4nlv	0	0.9167	1	0.9167	0.6667	1	1	0.8333	0.75	1	0.9042
qwen3.5-4b	Q4_K_M	mixed_q8k_q4nlv	1	0.9167	1	0.9167	0.6667	0.8	1	0.8021	0.75	1	0.8779
qwen3.5-4b	Q4_K_M	mixed_q8k_q4nlv	2	0.9167	1	0.8333	0.8333	0.8	1	0.8333	0.875	1	0.8967
qwen3.5-4b	Q4_K_M	mixed_q8k_q4nlv	3	0.8333	1	0.6333	0.6667	0.9	1	0.8333	0.75	1	0.8533
qwen3.5-4b	Q4_K_M	mixed_q8k_q4nlv	4	0.9167	1	0.7333	0.8333	0.8	1	0.8056	0.75	1	0.8686
qwen3.5-4b	Q4_K_M	q8	0	0.8333	1	0.9167	0.6667	1	1	0.6667	0.875	1	0.8708
qwen3.5-4b	Q4_K_M	q8	1	0.9167	1	0.8167	0.6667	0.9	1	0.8333	0.875	1	0.8967
qwen3.5-4b	Q4_K_M	q8	2	0.75	1	0.7333	0.8333	0.8	1	0.8021	0.875	1	0.8554
qwen3.5-4b	Q4_K_M	q8	3	0.8333	1	0.9167	0.6667	0.8	1	0.8333	0.875	1	0.8842
qwen3.5-4b	Q4_K_M	q8	4	0.8333	1	0.8167	0.6667	0.8	1	0.8333	0.75	1	0.8617
qwen3.5-4b	Q6_K	mixed_q8k_q4nlv	0	0.9167	1	0.9167	0.3333	0.9	1	0.8333	0.75	1	0.8775
qwen3.5-4b	Q6_K	mixed_q8k_q4nlv	1	0.9167	1	0.6333	0.1667	1	1	0.8333	0.875	1	0.8633
qwen3.5-4b	Q6_K	mixed_q8k_q4nlv	2	0.9167	1	0.7333	0.6667	0.8	1	0.8021	0.625	1	0.8471
qwen3.5-4b	Q6_K	mixed_q8k_q4nlv	3	0.75	1	0.9167	0.3333	0.9	1	0.8333	0.75	1	0.8525
qwen3.5-4b	Q6_K	mixed_q8k_q4nlv	4	0.8333	1	0.9167	0.6667	0.8	1	0.8333	0.75	1	0.8717
qwen3.5-4b	Q6_K	q8	0	0.8333	1	0.8167	0.6667	0.9	1	0.8333	0.875	1	0.8842
qwen3.5-4b	Q6_K	q8	1	0.9167	1	0.8167	0.6667	1	1	0.8333	0.75	1	0.8942

MODEL	QUANTIZATION	KV CACHE CONFIG	MTP DRAFT TOKENS	FINANCE	HR	IT	CUSTOMER SUPPORT	LEGAL	SALES	OPERATIONS	DATA ANALYTICS	KNOWLEDGE MANAGEMENT	OVERALL
qwen3.5-4b	Q6_K	q8	2	0.8333	1	0.9167	0.5	0.9	1	0.8333	0.75	1	0.8733
qwen3.5-4b	Q6_K	q8	3	0.8333	1	0.8167	0.8333	0.8	1	0.8333	0.875	1	0.8825
qwen3.5-4b	Q6_K	q8	4	0.9167	1	0.8167	0.5	0.8	1	0.8333	0.875	1	0.8783
qwen3.5-4b	Q8_0	mixed_q8k_q4nlv	0	0.9167	1	0.9167	0.8333	0.9	1	0.8333	0.875	1	0.915
qwen3.5-4b	Q8_0	mixed_q8k_q4nlv	1	0.9167	1	0.9167	0.8333	1	1	0.8333	0.75	1	0.9125
qwen3.5-4b	Q8_0	mixed_q8k_q4nlv	2	0.9167	1	0.8333	0.3333	0.9	1	0.8333	0.875	1	0.8817
qwen3.5-4b	Q8_0	mixed_q8k_q4nlv	3	0.8333	1	0.8167	0.5	1	1	0.8021	0.875	1	0.8796
qwen3.5-4b	Q8_0	mixed_q8k_q4nlv	4	0.75	1	0.8167	0.8333	0.9	1	0.8333	0.75	1	0.8675
qwen3.5-4b	Q8_0	q8	0	0.8333	1	0.6333	0.3333	0.9	1	0.8333	0.875	1	0.8492
qwen3.5-4b	Q8_0	q8	1	0.75	1	0.9167	0.6667	1	1	0.8333	0.875	1	0.8917
qwen3.5-4b	Q8_0	q8	2	0.8333	1	0.8333	0.8333	1	1	0.8333	0.75	1	0.8917
qwen3.5-4b	Q8_0	q8	3	0.8333	1	0.8167	0.6667	1	1	0.8333	0.625	1	0.8692
qwen3.5-4b	Q8_0	q8	4	0.75	1	1	0.8333	0.9	1	0.7708	0.75	1	0.8733
qwen3.5-9b	Q4_K_M	mixed_q8k_q4nlv	0	0.8333	1	0.9	0.8333	0.9	1	0.8333	0.875	1	0.9008
qwen3.5-9b	Q4_K_M	mixed_q8k_q4nlv	1	0.8333	1	0.9167	0.6667	1	1	0.75	0.875	0.875	0.875
qwen3.5-9b	Q4_K_M	mixed_q8k_q4nlv	2	0.8333	1	0.6	0.8333	0.8	1	0.75	0.75	1	0.8317
qwen3.5-9b	Q4_K_M	mixed_q8k_q4nlv	3	0.8333	1	0.6	0.6667	0.9	1	0.8333	0.75	0.875	0.8375
qwen3.5-9b	Q4_K_M	mixed_q8k_q4nlv	4	0.8333	1	1	0.6667	1	1	0.75	0.875	1	0.8958
qwen3.5-9b	Q4_K_M	q8	0	0.8333	1	1	0.8333	0.7	1	0.8333	0.75	1	0.8783
qwen3.5-9b	Q4_K_M	q8	1	0.75	1	0.8167	0.6667	0.9	1	0.8333	0.75	1	0.8592
qwen3.5-9b	Q4_K_M	q8	2	0.75	1	0.9	0.6667	1	1	0.75	1	1	0.8858
qwen3.5-9b	Q4_K_M	q8	3	0.9167	1	1	0.6667	0.9	1	0.75	0.75	0.875	0.8733
qwen3.5-9b	Q4_K_M	q8	4	0.8333	1	0.8167	0.6667	1	1	0.8333	0.875	1	0.8942
qwen3.5-9b	Q6_K	mixed_q8k_q4nlv	0	0.8333	1	0.9167	0.6667	1	1	0.75	0.75	1	0.875
qwen3.5-9b	Q6_K	mixed_q8k_q4nlv	1	0.8333	1	0.9167	0.6667	0.9	1	0.8333	0.875	1	0.8942
qwen3.5-9b	Q6_K	mixed_q8k_q4nlv	2	0.75	1	0.9	0.6667	1	1	0.8333	0.875	1	0.89
qwen3.5-9b	Q6_K	mixed_q8k_q4nlv	3	0.8333	1	1	0.8333	0.9	1	0.75	0.875	1	0.8942

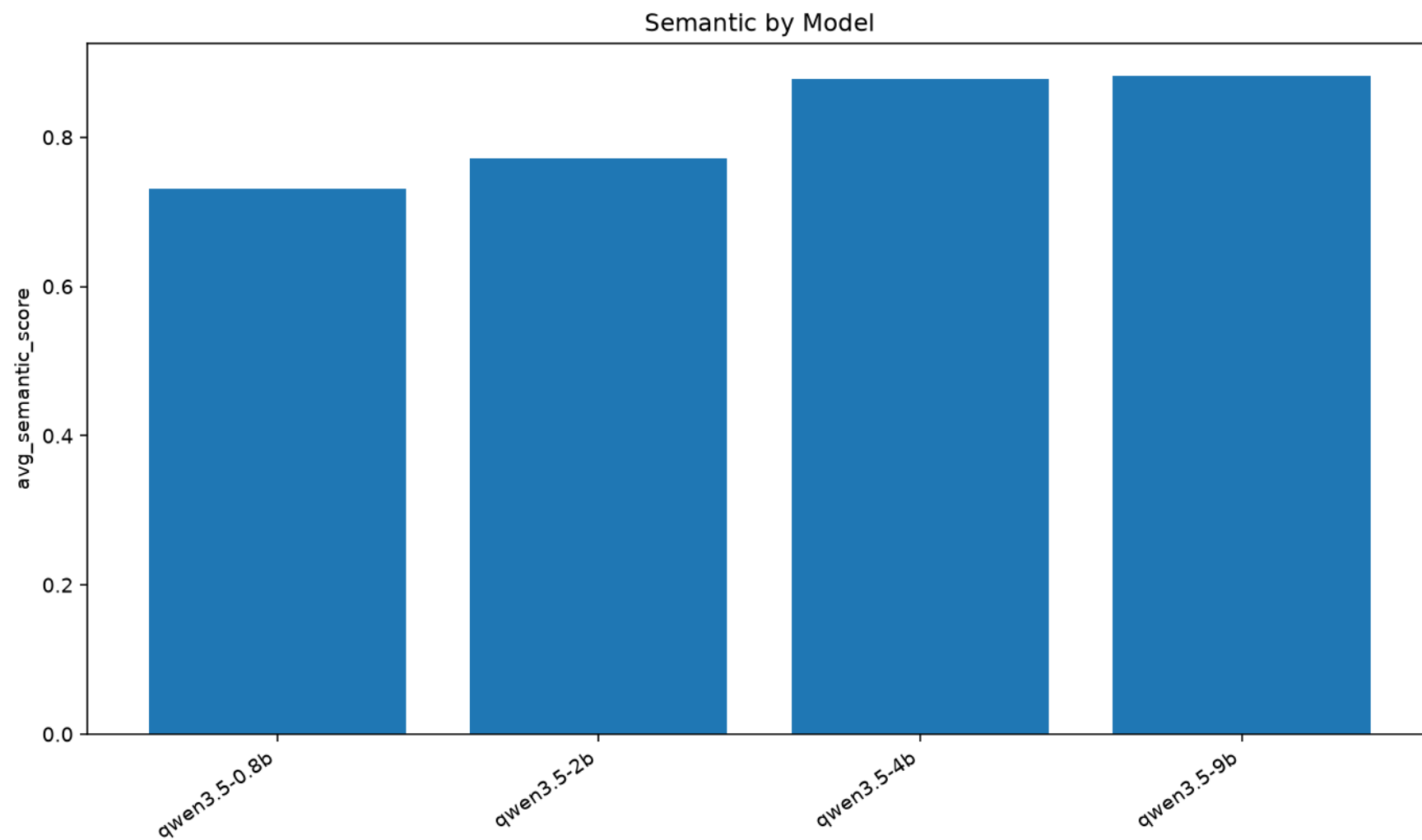
MODEL	QUANTIZATION	KV CACHE CONFIG	MTP DRAFT TOKENS	FINANCE	HR	IT	CUSTOMER SUPPORT	LEGAL	SALES	OPERATIONS	DATA ANALYTICS	KNOWLEDGE MANAGEMENT	OVERALL
qwen3.5-9b	Q6_K	mixed_q8k_q4nlv	4	0.8333	1	0.9167	0.5	0.9	1	0.75	0.875	1	0.8692
qwen3.5-9b	Q6_K	q8	0	0.8333	1	1	0.8333	1	1	0.8021	0.75	1	0.9021
qwen3.5-9b	Q6_K	q8	1	0.8333	1	0.9	0.6667	0.8	1	0.8333	0.75	1	0.87
qwen3.5-9b	Q6_K	q8	2	0.8333	1	1	0.6667	0.9	1	0.8333	0.875	1	0.9025
qwen3.5-9b	Q6_K	q8	3	0.8333	1	0.9	0.6667	0.9	1	0.75	0.875	1	0.8758
qwen3.5-9b	Q6_K	q8	4	0.8333	1	0.8167	0.8333	0.9	1	0.8333	0.75	1	0.88
qwen3.5-9b	Q8_0	mixed_q8k_q4nlv	0	0.8333	1	1	0.6667	0.8	1	0.75	0.75	1	0.8633
qwen3.5-9b	Q8_0	mixed_q8k_q4nlv	1	0.8333	1	0.9167	0.8333	0.9	1	0.8333	0.875	1	0.9025
qwen3.5-9b	Q8_0	mixed_q8k_q4nlv	2	0.8333	1	0.8167	0.8333	1	1	0.8333	0.75	1	0.89
qwen3.5-9b	Q8_0	mixed_q8k_q4nlv	3	0.9167	1	1	0.6667	1	1	0.8333	0.75	0.875	0.9
qwen3.5-9b	Q8_0	mixed_q8k_q4nlv	4	0.8333	1	0.5167	0.8333	1	1	0.75	0.75	1	0.8433
qwen3.5-9b	Q8_0	q8	0	0.8333	1	0.9167	0.6667	1	1	0.8333	0.875	1	0.9042
qwen3.5-9b	Q8_0	q8	1	0.8333	1	0.9167	0.8333	1	1	0.75	0.875	1	0.8958
qwen3.5-9b	Q8_0	q8	2	0.8333	1	0.9167	0.8333	1	1	0.75	0.875	1	0.8958
qwen3.5-9b	Q8_0	q8	3	0.8333	1	0.9167	0.8333	0.8	1	0.75	0.75	1	0.8633
qwen3.5-9b	Q8_0	q8	4	0.8333	1	0.9167	0.8333	1	1	0.8333	0.875	1	0.9125

Pareto Configurations

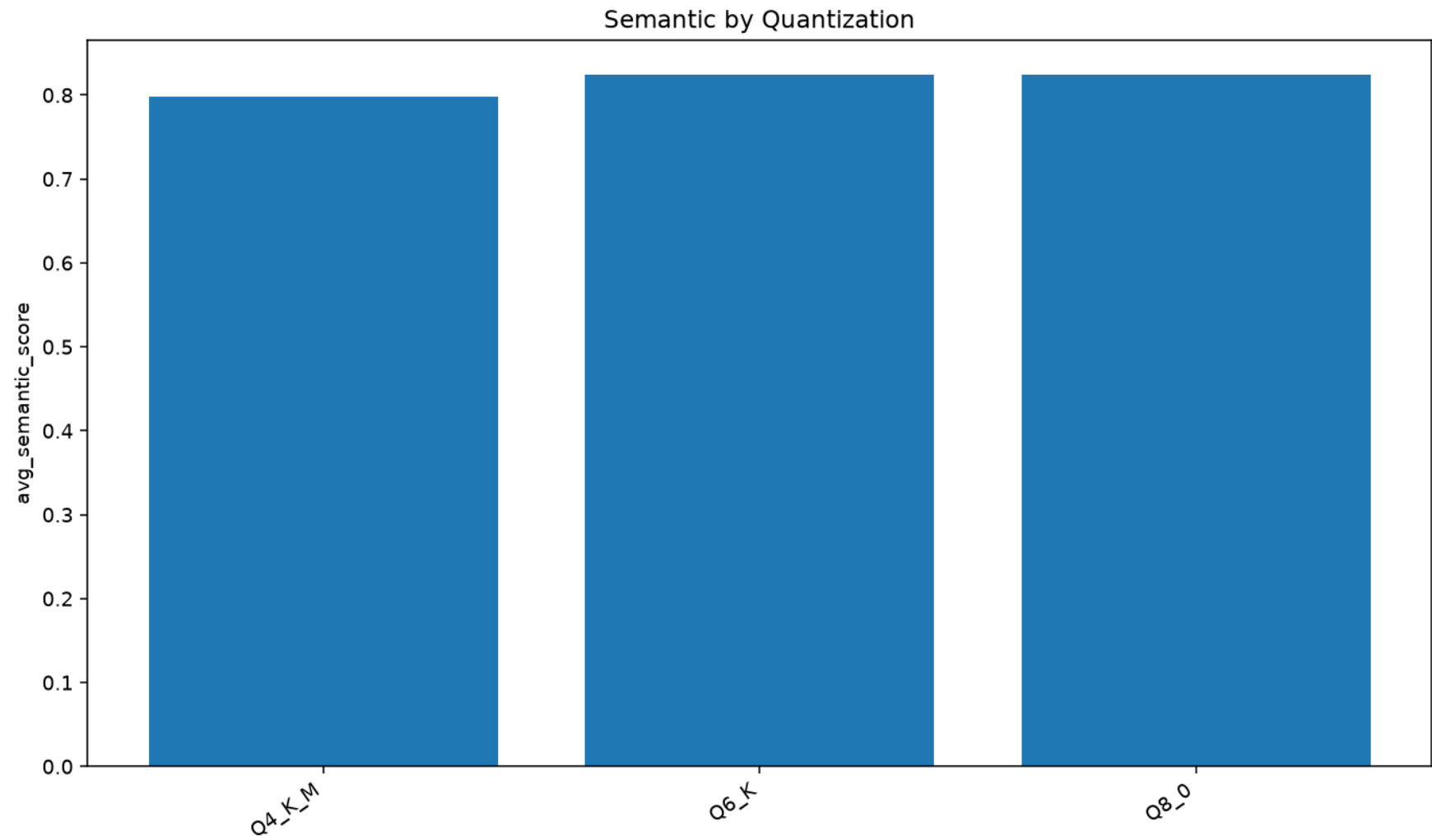
MODEL	QUANTIZATION	KV CACHE CONFIG	MTP DRAFT TOKENS	RUNS	AVG SEMANTIC SCORE	OVERALL PASS RATE	AVG ELAPSED S	AVG GENERATION TOK S	AVG PROMPT TOK S
qwen3.5-0.8b	Q4_K_M	mixed_q8k_q4nlv	0	20	0.8133	0.7	0.8791	176.6	3039.9
qwen3.5-0.8b	Q4_K_M	q8	0	20	0.6929	0.4	0.691	216.0	4405.9
qwen3.5-0.8b	Q6_K	mixed_q8k_q4nlv	4	20	0.8386	0.7	1.354	100.6	2619.8
qwen3.5-0.8b	Q6_K	q8	0	20	0.7538	0.55	0.6655	205.5	4717.6
qwen3.5-0.8b	Q6_K	q8	1	20	0.7675	0.65	0.6999	186.3	4020.5
qwen3.5-0.8b	Q6_K	q8	2	20	0.7858	0.6	0.7937	165.7	3953.0
qwen3.5-2b	Q8_0	mixed_q8k_q4nlv	2	20	0.82	0.65	1.3007	100.9	2311.8
qwen3.5-2b	Q8_0	q8	0	20	0.8233	0.7	1.3184	97.34	3450.1
qwen3.5-4b	Q4_K_M	mixed_q8k_q4nlv	0	20	0.9042	0.8	2.395	62.31	1306.3
qwen3.5-4b	Q4_K_M	mixed_q8k_q4nlv	2	20	0.8967	0.75	1.9135	86.40	1212.0
qwen3.5-4b	Q4_K_M	q8	1	20	0.8967	0.8	1.5285	88.39	1668.1
qwen3.5-4b	Q4_K_M	q8	2	20	0.8554	0.75	1.425	96.34	1617.1
qwen3.5-4b	Q4_K_M	q8	3	20	0.8842	0.75	1.4458	96.69	1653.7
qwen3.5-4b	Q8_0	mixed_q8k_q4nlv	0	20	0.915	0.85	2.7165	43.70	1183.4
qwen3.5-4b	Q8_0	mixed_q8k_q4nlv	1	20	0.9125	0.85	2.1977	60.55	1126.9

Charts

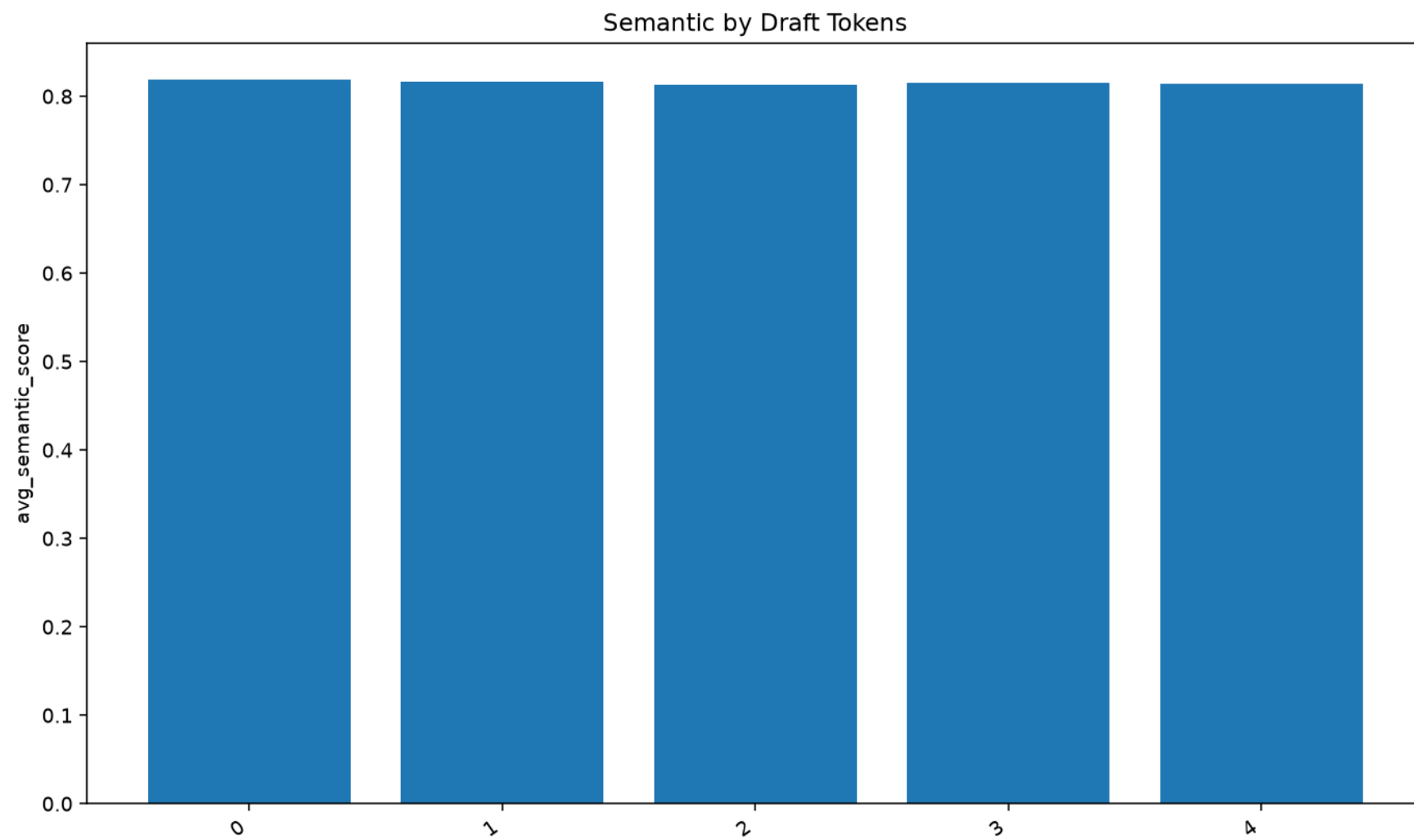
Average semantic score by model size



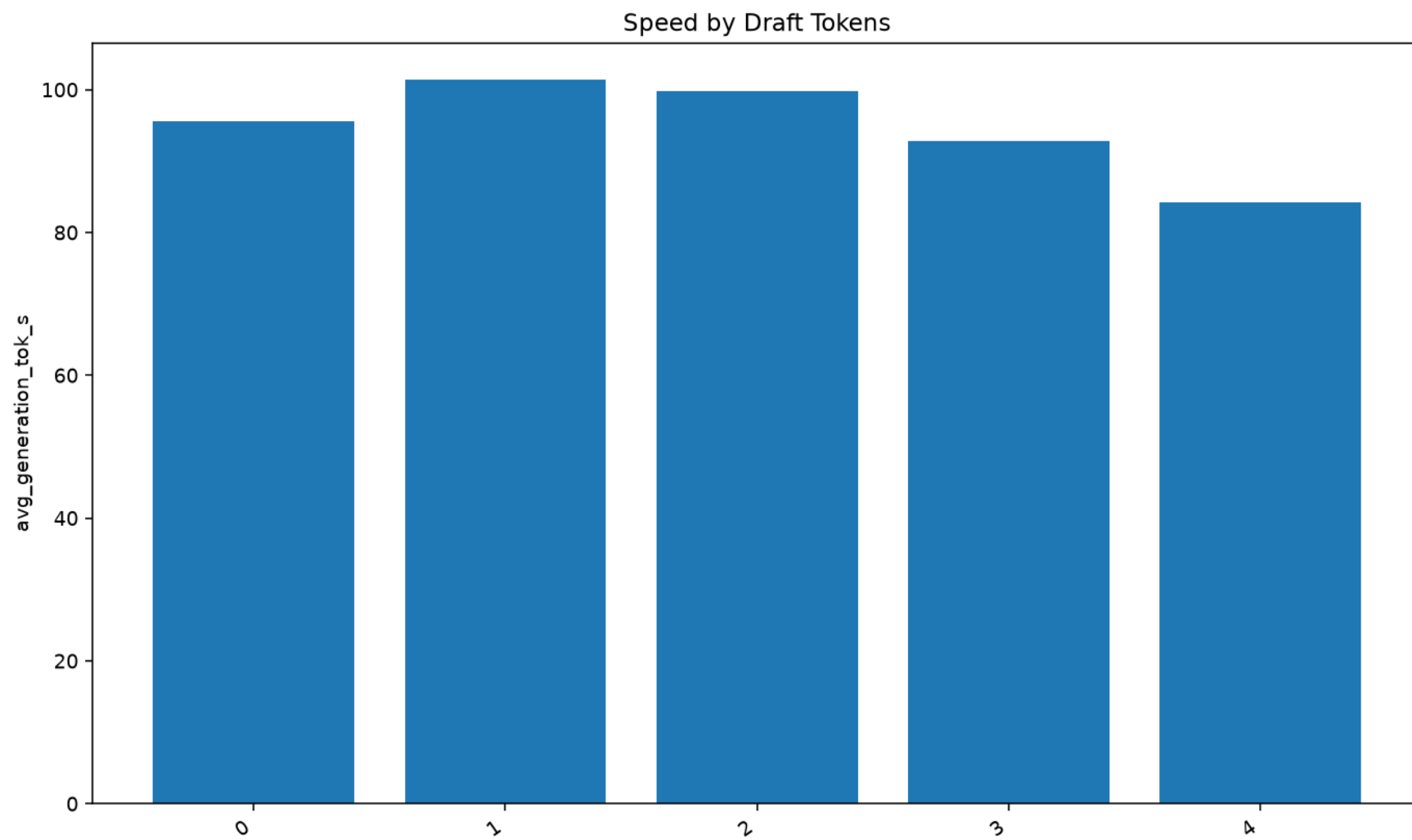
Semantic score by quantisation



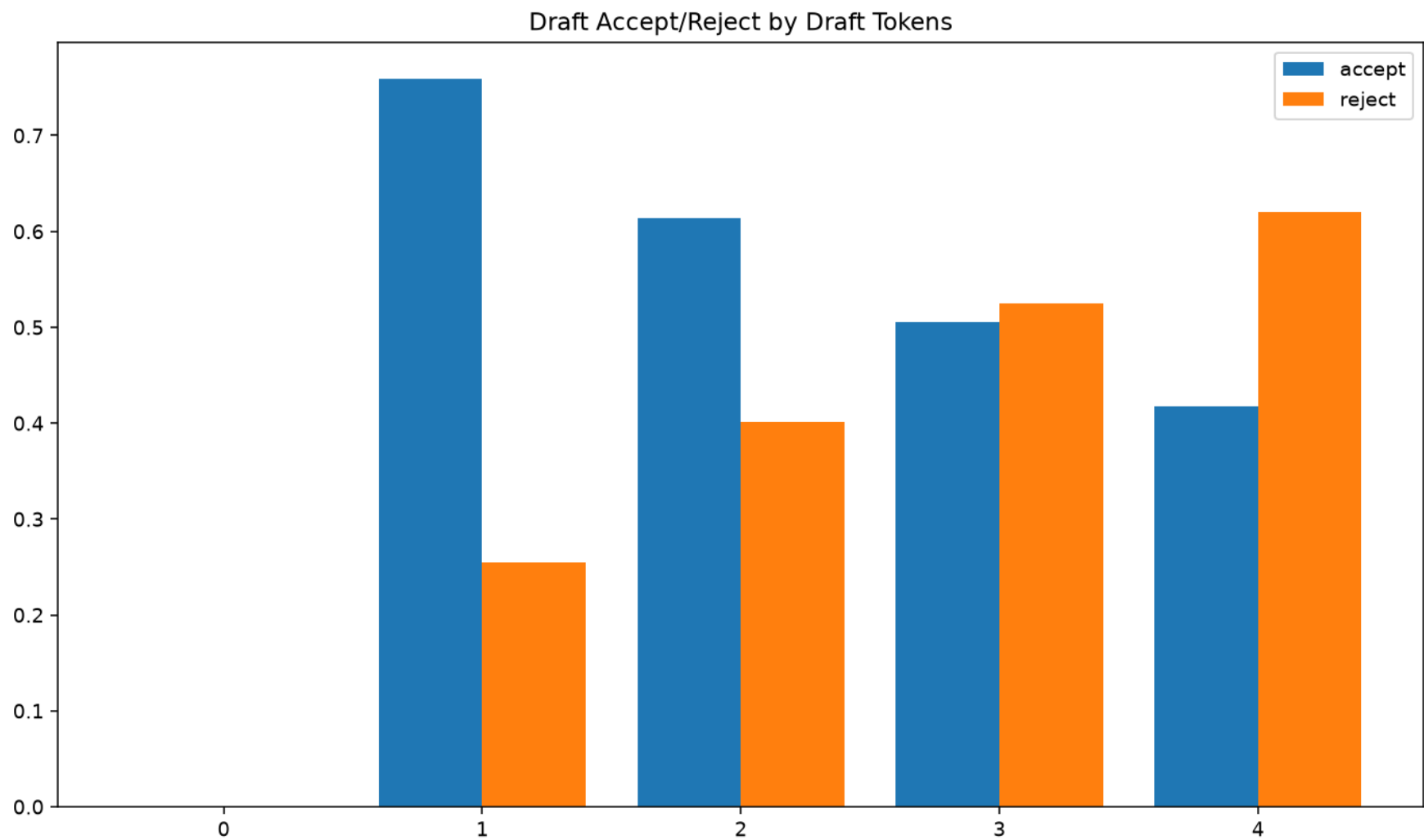
Semantic score by MTP draft-token setting



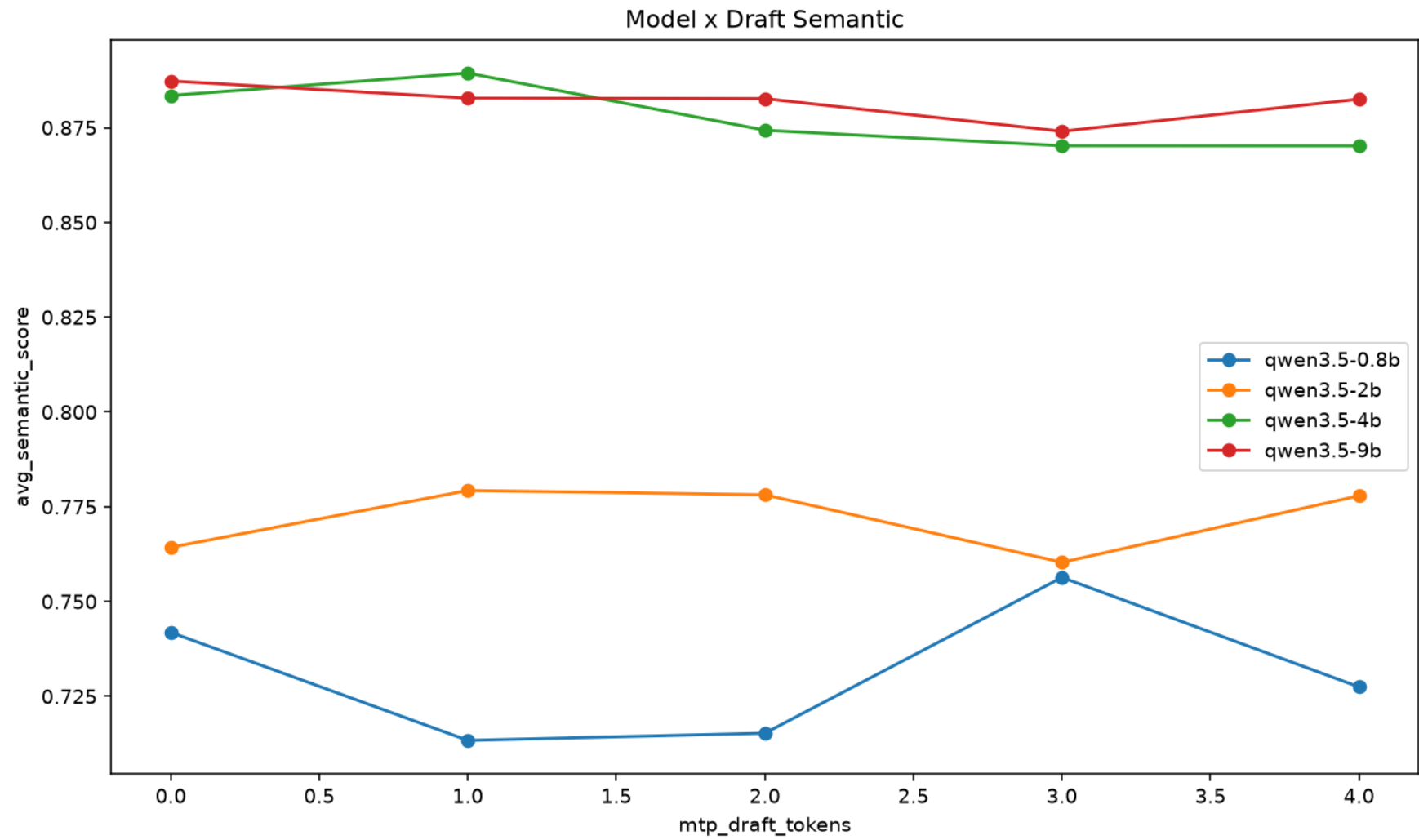
Generation speed by MTP draft-token setting



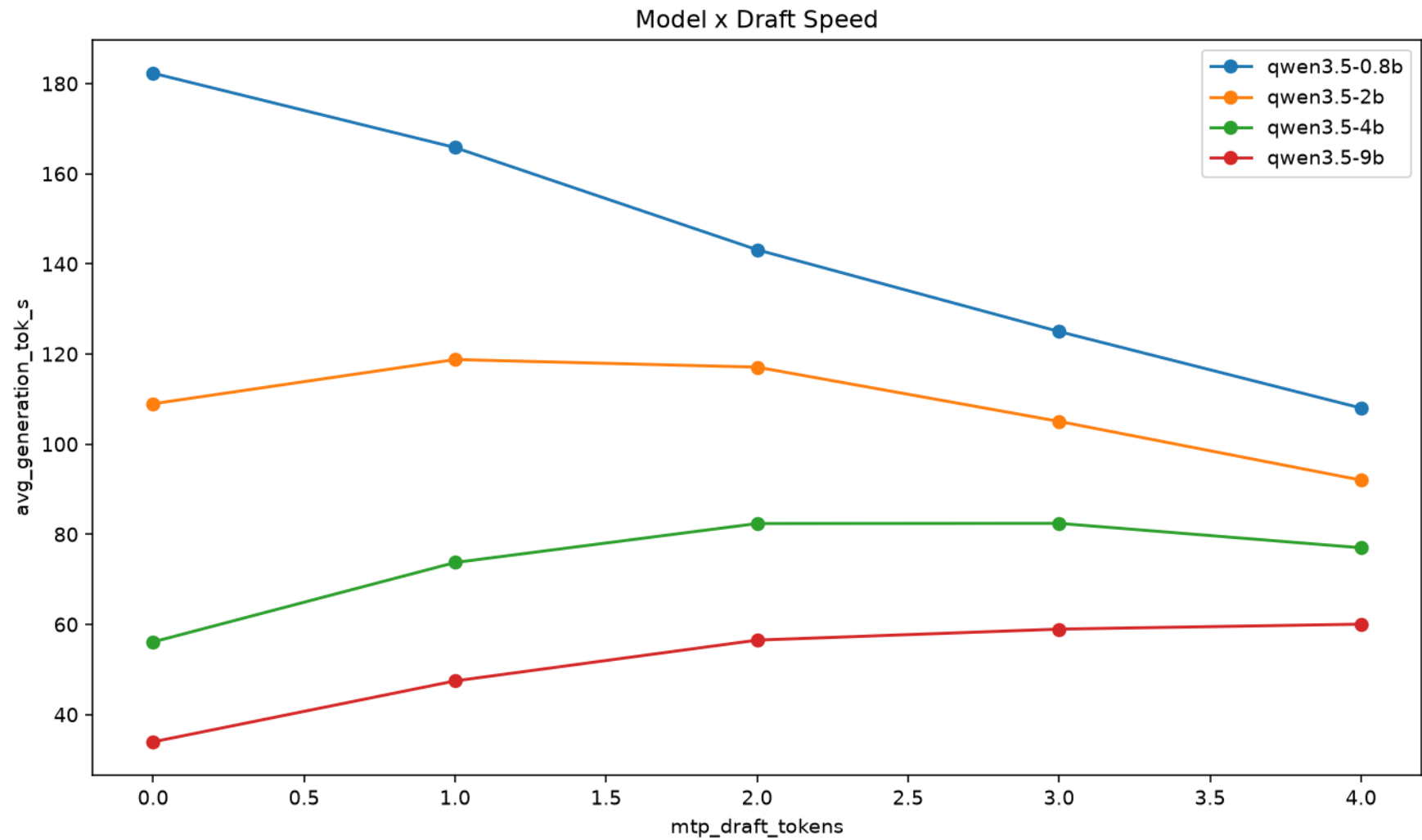
Draft-token accept/reject behaviour



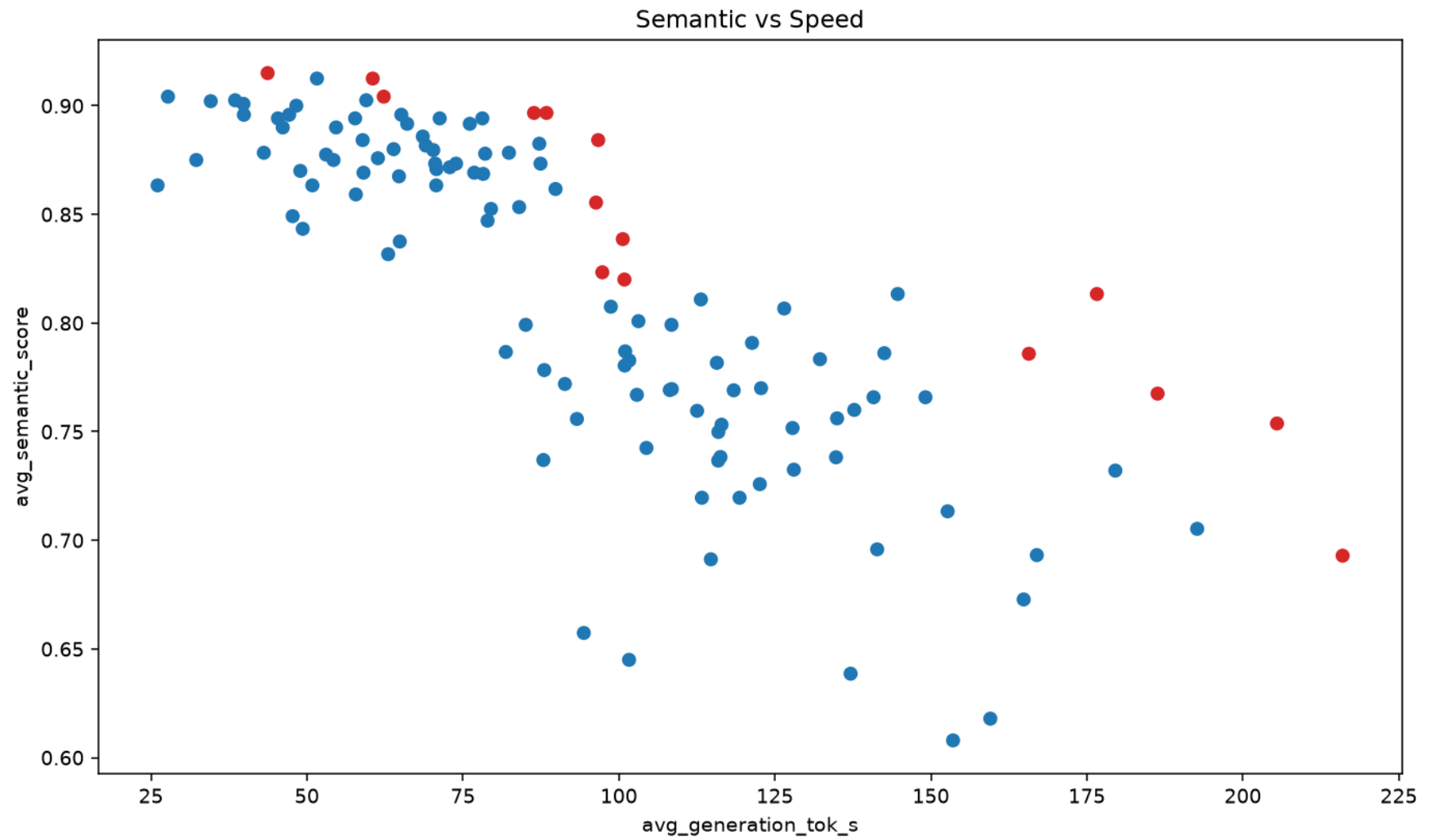
Model by draft-token interaction for semantic score



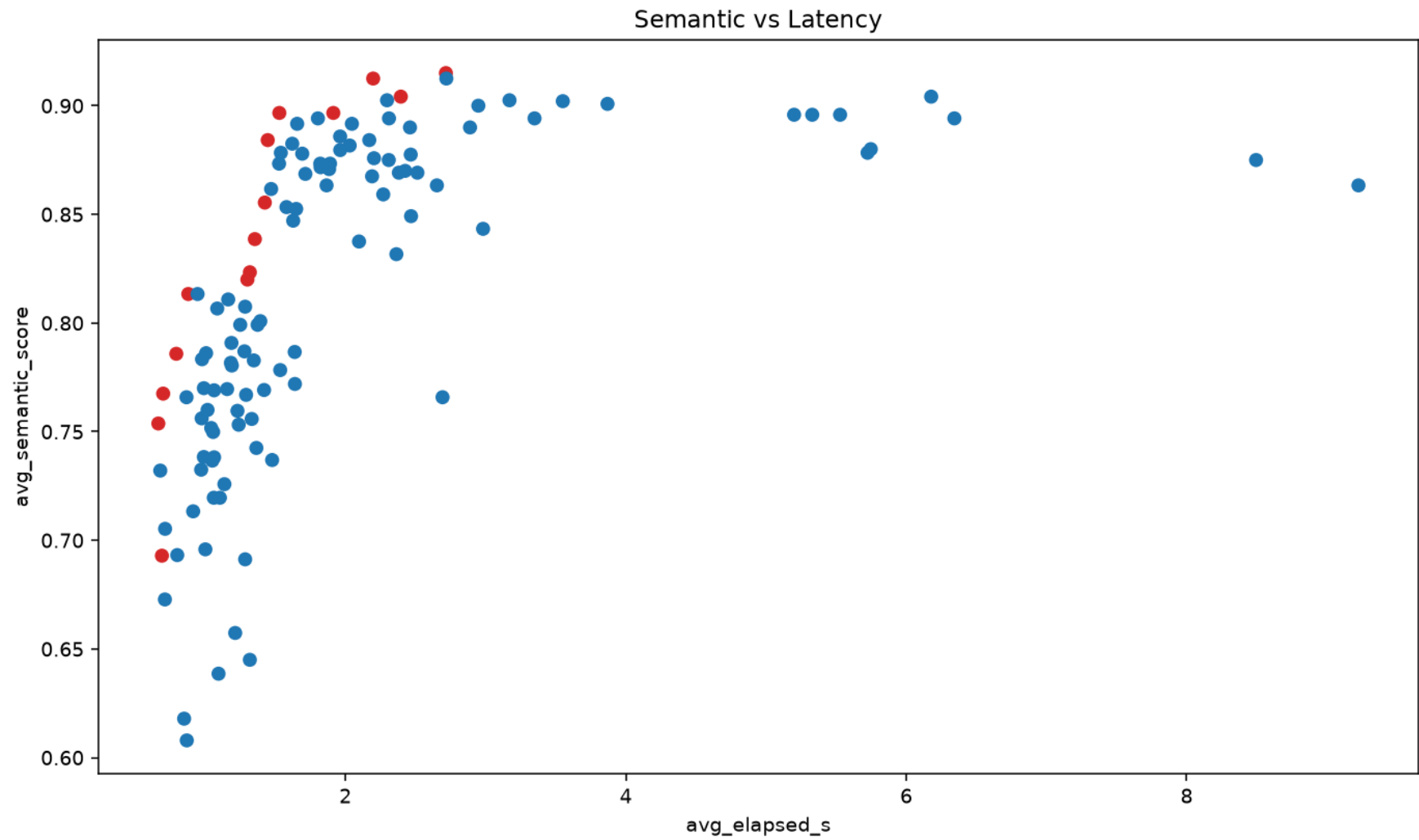
Model by draft-token interaction for speed



Pareto view: semantic score versus speed



Pareto view: semantic score versus latency



Enterprise function suitability heatmap



Limitations And Next Tests

This analysis uses deterministic benchmark scoring rubrics. Confidence intervals are bootstrap estimates over benchmark rows. ANOVA tables were generated through statsmodels OLS Type-II ANOVA.

The dataset was produced by merging the completed MTP 0/1 run with the later MTP 2/3/4 run.

The results are preliminary. Next tests should broaden context-size sweeps, long-context behaviour, workload-specific prompt families, MTP behaviour under long context, KV-cache quantisation interactions, and prompt-router design.