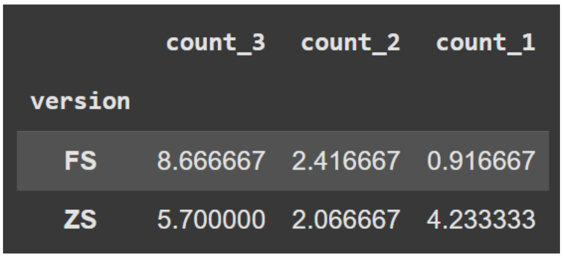
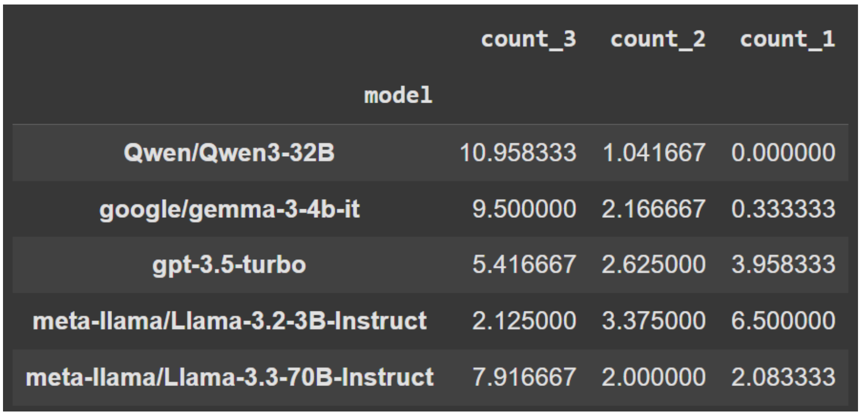
LLM-as-a-Judge Quality Summary



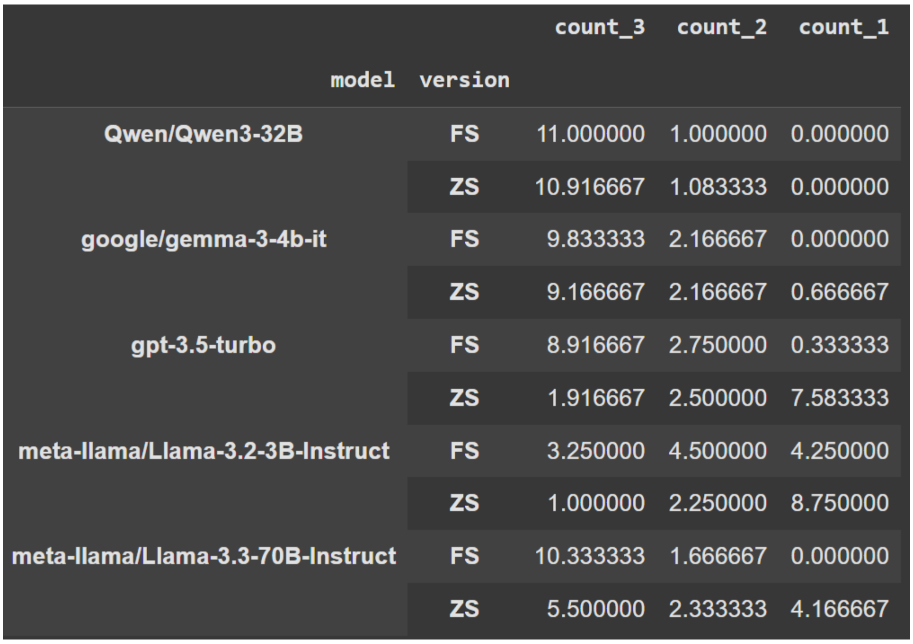
	count_3	count_2	count_1
version			
FS	8.666667	2.416667	0.916667
ZS	5.700000	2.066667	4.233333

Figure 1. Average quality item count per setting (Few-Shot and Zero-Shot).



	count_3	count_2	count_1
model			
Qwen/Qwen3-32B	10.958333	1.041667	0.000000
google/gemma-3-4b-it	9.500000	2.166667	0.333333
gpt-3.5-turbo	5.416667	2.625000	3.958333
meta-llama/Llama-3.2-3B-Instruct	2.125000	3.375000	6.500000
meta-llama/Llama-3.3-70B-Instruct	7.916667	2.000000	2.083333

Figure 2. Average quality item count per model.



		count_3	count_2	count_1
model	**version**			
Qwen/Qwen3-32B	**FS**	11.000000	1.000000	0.000000
	ZS	10.916667	1.083333	0.000000
google/gemma-3-4b-it	**FS**	9.833333	2.166667	0.000000
	ZS	9.166667	2.166667	0.666667
gpt-3.5-turbo	**FS**	8.916667	2.750000	0.333333
	ZS	1.916667	2.500000	7.583333
meta-llama/Llama-3.2-3B-Instruct	**FS**	3.250000	4.500000	4.250000
	ZS	1.000000	2.250000	8.750000
meta-llama/Llama-3.3-70B-Instruct	**FS**	10.333333	1.666667	0.000000
	ZS	5.500000	2.333333	4.166667

Figure 3. Quality item count per setting, per model.