Model Index	Model Ref	Model Name	Model Name	Version	Context Length	Min	Sec	Latency (s / record) [1]	Input Token Price ($ / 1K tokens)	Input Tokens	Output Token Price ($ / 1K tokens)	Output Tokens	Total Price ($ = Input + Output)	Avg Price ($ / record)	Num Records	TP	TN	FP	FN	AP	AN	Precision	Recall	F1	Num Manual Clean	Notes
00			Gemini 1.5 Flash	Original				--						--								0.3555	0.6646	0.4632		The first baseline produced on 12/13/2024 with the dataset.
01			Gemini 1.5 Flash	Baseline				--						--								0.6369	0.6646	0.6505		Improved from the 00 version by extracting better excerpts & revised prompt.
23	0.gpt4o_mini_baseline_prompt_eval.csv	gpt-4o-mini-2024-07-18	GPT 4o Mini	Baseline	128k	11	14	0.2057	$0.0002	2,065,575	$0.0006	173,604	$0.41	$0.00	1084	126	595	340	23	149	935	0.2704	0.8456	0.4098	0	
24	1.gpt4o_mini_v213_prompt_eval.csv	gpt-4o-mini-2024-07-18	GPT 4o Mini	v213 prompt	128k	6	35	0.1375	$0.0002	3,736,019	$0.0006	86,503	$0.61	$0.00	1084	115	755	180	34	149	935	0.3898	0.7718	0.5180	0	
25	2.gpt4o_mini_v214_v2_prompt_eval.csv	gpt-4o-mini-2024-07-18	GPT 4o Mini	v214_v2 prompt	128k	6	32	0.1347	$0.0002	3,705,667	$0.0006	87,667	$0.61	$0.00	1084	117	743	192	32	149	935	0.3786	0.7852	0.5109	0	
26	3.gpt4o_mini_v217_prompt_eval.csv	gpt-4o-mini-2024-07-18	GPT 4o Mini	v217 prompt	128k	6	52	0.1531	$0.0002	3,626,535	$0.0006	82,471	$0.59	$0.00	1084	114	759	176	35	149	935	0.3931	0.7651	0.5194	0	
27	4.gpt4o_v213_prompt_eval.csv	gpt-4o-2024-11-20	GPT 4o	v213 prompt	128k	10	25	0.1983	$0.0025	3,736,019	$0.0100	109,085	$10.43	$0.01	1084	120	873	62	29	149	935	0.6593	0.8054	0.7251	0	
28	5.gpt4o_v217_prompt_eval.csv	gpt-4o-2024-11-20	GPT 4o	v217 prompt	128k	11	43	0.2325	$0.0025	3,626,535	$0.0100	112,956	$10.20	$0.01	1084	117	900	35	32	149	935	0.7697	0.7852	0.7774	0	
29	6.o1_mini_baseline_prompt_eval.csv	o1-mini-2024-09-12	o1 Mini	Baseline prompt	128k	24	30	0.4483	$0.0011	2,101,807	$0.0044	1,639,308	$9.52	$0.01	1084	106	818	117	43	149	935	0.4753	0.7114	0.5699	5	
30	7.o1_mini_v213_prompt_eval.csv	o1-mini-2024-09-12	o1 Mini	v213 prompt	128k	29	9	0.5166	$0.0011	2,095,535	$0.0044	1,855,244	$10.47	$0.01	1084	93	844	91	56	149	935	0.5054	0.6242	0.5586	0	
31	8.o1_mini_v217_prompt_eval.csv	o1-mini-2024-09-12	o1 Mini	v217 prompt	128k	27	27	0.4982	$0.0011	3,125,103	$0.0044	2,142,197	$12.86	$0.01	1084	100	816	119	49	149	935	0.4566	0.6711	0.5435	5	
32	9.o3_mini_v213_prompt_eval.csv	o3-mini-2025-01-31	o3 Mini	v213 prompt	200k	23	13	0.4151	$0.0011	2,074,247	$0.0044	1,465,726	$8.73	$0.01	1084	89	889	46	60	149	935	0.6593	0.5973	0.6268	0	
33	10.o3_mini_v217_prompt_eval.csv	o3-mini-2025-01-31	o3 Mini	v217 prompt	200k	39	45	0.7251	$0.0011	3,041,175	$0.0044	2,208,300	$13.06	$0.01	1084	97	880	55	52	149	935	0.6382	0.6510	0.6445	0	

Note:

1. all models use the same dataset. Other than the Original version, all models use the same excerpts as the context for the classification.
2. parameters for experiments: model provider/version & prompt.

[1] ThreadPool on 8 Core Apple M2 with 5 max workers