

Queries Per Model	Method	Qwen-3B		Qwen-7B		Mistral-7B		Llama-8B	
		Rank ↓	Acc. ↑	Rank ↓	Acc. ↑	Rank ↓	Acc. ↑	Rank ↓	Acc. ↑
-	Random Selection	233.8	0.588	318.4	0.599	144.6	0.431	270.8	0.601
	Best Base	56.5	0.716	30.0	0.783	62.0	0.556	41.0	0.713
10	Uniform	166.3	0.671	173.8	0.724	29.5	0.656	222.0	0.665
	UCB	88.0	0.708	83.9	0.770	14.1	0.684	110.2	0.702
	UCB-StdDev.	81.8	0.710	75.2	0.772	11.8	0.686	96.0	0.707
	TTTS	92.6	0.708	75.7	0.772	14.0	0.685	103.6	0.702
	Successive Rejects	94.4	0.706	128.3	0.752	25.6	0.662	171.5	0.683
	BayesElim	56.4	0.717	58.9	0.778	13.5	0.685	79.9	0.708
	UCB-E	78.4	0.710	83.7	0.769	13.6	0.684	106.5	0.703
	Sequential Halving	69.3	0.714	62.9	0.777	15.2	0.684	75.2	0.710
	Ours	11.3	0.726	15.8	0.786	2.6	0.694	15.4	0.725
50	Uniform	91.2	0.707	81.9	0.770	17.4	0.683	114.9	0.705
	UCB	41.2	0.719	39.8	0.782	8.9	0.689	55.8	0.718
	UCB-StdDev.	34.7	0.721	28.9	0.784	4.0	0.693	37.1	0.722
	TTTS	53.1	0.717	51.7	0.780	10.8	0.687	70.6	0.712
	Successive Rejects	83.1	0.710	78.4	0.773	16.5	0.682	92.0	0.707
	BayesElim	30.0	0.721	31.0	0.784	6.7	0.691	34.4	0.721
	UCB-E	37.4	0.720	33.1	0.783	4.3	0.693	43.4	0.720
	Sequential Halving	41.0	0.719	29.6	0.784	7.7	0.690	29.9	0.720
	Ours	3.5	0.729	3.6	0.790	1.6	0.695	3.0	0.736
	Best Available	1.0	0.732	1.0	0.791	1.0	0.696	1.0	0.747