Table 1: Alternative methods for navigating logic blocks at inference time without RL. The **bold** numbers indicate the best performance.

LLM	Method	Task			Average
		GSM8K	GPQA	StrategyQA	
Qwen2.5-Instruct-7B	DirectQA	91.58	31.25	68.85	63.89
	Fixed logic sequence	88.38	36.16	71.13	65.22
	Supervise-trained navigator	89.84	36.83	70.31	65.66
	LLM as navigator	89.16	37.95	74.38	67.16
	RL-trained navigator (ours)	**92.87**	**44.64**	**79.04**	**72.18**

Table 2: Alternative methods for generating reward signals without relying on a pre-trained PRM. The **bold** numbers indicate the best performance.

LLM	Method	Task			Average
		GSM8K	GPQA	StrategyQA	
Qwen2.5-Instruct-7B	DirectQA	91.58	31.25	68.85	63.89
	LLM as navigator	89.16	37.95	74.38	67.16
	RLoT with ORM	91.96	41.52	73.94	69.14
	RLoT with PRM (ours)	**92.87**	**44.64**	**79.04**	**72.18**

Table 3: An example of self-assessing of intermediate states.

Question	Reasoning step	Self-evaluation score	
		Correctness of modeling	Correctness of calculation
Mark has a garden with flowers. He planted plants of three different colors in it. Ten of them are yellow, and there are 80% more of those in purple. There are only 25% as many green flowers as there are yellow and purple flowers. How many flowers does Mark have in his garden?	In this step, we aim at calculate the number of purple flowers, which is 80% more than the yellow ones. We calculate by 10 \times (1 + 0.8) = 18	True	True
	In this step, we aim at calculate the number of purple flowers, which is 80% more than the yellow ones. We calculate by 10 \times 0.8 = 8	False	True
	In this step, we aim at calculate the number of purple flowers, which is 80% more than the yellow ones. We calculate by 10 \times (1 + 0.8) = 17	True	False

Table 4: The performance of RLoT with smaller LLMs. The **bold** numbers indicate our method.

LLM	Size	Method	Task			Average	Gap
			GSM8K	GPQA	StrategyQA		
Qwen2.5-Instruct	3B	Few-shot CoT	79.91	31.47	66.38	59.25	-17.98
		RLoT (ours)	**81.57**	**43.53**	**72.78**	**65.96**	**-11.27**
	7B	Few-shot CoT	91.60	36.40	74.38	67.46	-9.77
		RLoT (ours)	**92.87**	**44.64**	**79.04**	**72.18**	**-5.05**
	14B	Few-shot CoT	94.80	45.50	78.60	72.97	-4.27
		RLoT (ours)	**94.16**	**51.34**	**81.22**	**75.57**	**-1.66**
	72B	Few-shot CoT	95.80	49.00	86.90	77.23	–

Table 5: Token consumption per question across tasks. The **bold** numbers indicate our method.

Model	Method	MATH		GSM8K		GPQA		MMLU-STEM		StrategyQA		Average	
		Input	Output	Input	Output	Input	Output	Input	Output	Input	Output	Input	Output
Qwen2.5-14B-Instruct	Direct QA	46	362	55	183	159	460	82	210	30	259	74	295
	Zero-shot CoT	58	367	66	206	175	546	87	334	33	376	84	366
	Few-shot CoT	466	336	1089	135	1557	589	1114	259	421	223	929	308
	CoT-SC	2014	1320	4339	543	6295	1945	4448	1039	1586	897	3736	1149
	ToT	4983	6277	3570	4312	6063	7404	4797	5363	5762	6491	5035	5969
	RLoT (ours)	**3735**	**1109**	**2310**	**634**	**5501**	**1485**	**2923**	**791**	**2348**	**756**	**3363**	**955**
Qwen2.5-7B-Instruct	Direct QA	44	366	58	182	159	546	66	404	22	279	70	355
	Zero-shot CoT	47	393	78	222	184	522	82	413	32	401	85	390
	Few-shot CoT	483	581	1074	188	1556	442	1128	221	396	348	927	356
	CoT-SC	1932	1443	4379	750	6290	1800	4433	883	1612	1391	3729	1253
	ToT	6679	7525	5792	5301	5771	6489	5058	4669	3807	2013	5421	5199
	RLoT (ours)	**1437**	**807**	**1283**	**476**	**2112**	**889**	**1028**	**434**	**1735**	**606**	**1519**	**643**
Llama3.1-8B-Instruct	Direct QA	44	437	53	180	174	811	71	406	13	201	71	407
	Zero-shot CoT	73	495	64	215	175	661	91	365	20	418	85	431
	Few-shot CoT	510	472	1101	173	1568	769	1127	168	417	272	945	371
	CoT-SC	2049	1885	4325	808	6120	3078	4429	808	1582	1071	3701	1530
	ToT	3474	4307	5212	4780	5615	3794	5331	5999	4320	4868	4790	4750
	RLoT (ours)	**4481**	**1981**	**2842**	**1028**	**9818**	**3387**	**4408**	**1486**	**3049**	**1094**	**4920**	**1795**
GPT-4o-mini	Direct QA	43	404	54	223	165	397	86	175	13	201	72	280
	Zero-shot CoT	50	422	68	260	160	525	89	326	16	382	77	383
	Few-shot CoT	513	408	1107	205	1575	505	1119	275	409	254	945	329
	CoT-SC	2030	1554	4442	823	6305	2027	4360	1005	1601	1014	3748	1285
	ToT	4534	5649	5890	5810	5175	5841	4950	5324	3833	1467	4876	4818
	RLoT (ours)	**3483**	**1575**	**3069**	**1048**	**4372**	**1550**	**1882**	**710**	**1612**	**673**	**2884**	**1111**
Average	Direct QA	44	392	55	192	164	554	76	299	20	235	72	334
	Zero-shot CoT	57	419	69	226	174	564	87	360	25	394	82	392
	Few-shot CoT	493	449	1093	175	1564	576	1122	231	411	274	937	341
	CoT-SC	2006	1551	4371	731	6253	2213	4418	934	1595	1093	3729	1304
	ToT	4918	5940	5116	5051	5656	5882	5034	5339	4431	3710	5031	5184
	RLoT (ours)	**3284**	**1368**	**2376**	**797**	**5451**	**1828**	**2560**	**855**	**2186**	**782**	**3171**	**1126**
RLoT plus training cost		**3638**	**1470**	**2730**	**899**	**5804**	**1930**	**2914**	**958**	**2540**	**885**	**3525**	**1228**

Table 6: Extended baseline comparisons. The **bold** numbers indicate the best performance in each group of experiments, and the underlined numbers indicate the best baseline method.

LLM	Method		Task			Average
			GSM8K	GPQA	StrategyQA	
Qwen2.5-Instruct-7B	DirectQA		91.58	31.25	68.85	63.89
	DeAR		88.38	<u>36.16</u>	71.13	65.22
	Refine		88.55	32.59	73.94	65.03
	Tree search	One-step-greedy	89.08	32.59	73.07	64.91
		MCTS	90.07	34.82	75.69	66.86
		Litesearch	89.76	33.04	75.84	66.21
		Q*	91.51	34.60	<u>77.00</u>	<u>67.70</u>
	Graph of Thoughts (GoT)		88.96	33.04	75.42	65.81
	Buffer of Thoughts (BoT)		<u>92.35</u>	34.51	74.33	67.06
	RLoT (ours)		**92.87**	**44.64**	**79.04**	**72.18**

Table 7: Extended results for the baselines mentioned in the Appendix (in progress). The **bold** numbers indicate the best performance in each category.

LLM	Category	Method	Task					Average
			MATH	GSM8K	GPQA	MMLU-STEM	StrategyQA	
Llama3.1-8B-Insturct	Finetuning	AceMath	**64.42**	**90.45**		**75.30**		
		PFPO	57.80	89.60				
	Inference time	LLM2	48.60	88.00				
		LLaMA-Berry	54.80	89.80		68.30		
		HiAR-ICL	55.00	**90.70**			73.20	
		RLoT (ours)	**56.56**	90.07	**46.88**	**80.56**	**84.42**	**71.70**

Table 8: Results on directly using PRM at the inference time and keeping other part of the inference pipeline the same.

LLM	Navigator	Task	Accuracy	LLM consumption		Navigator consumption
				Input (token)	Output (token)	Inference time (s)
Qwen2.5-Instruct-7B	PRM directly	GSM8K	89.08	4145	1448	7.91
		GPQA	32.59	13521	4360	12.37
		StrategyQA	73.07	4351	1796	8.23
		Average	64.91	7339	2535	9.50
	RLoT (ours)	GSM8K	92.87	5287	1569	<0.01
		GPQA	44.64	7345	1733	<0.01
		StrategyQA	79.04	5864	1760	<0.01
		Average	72.18	6165	1687	<0.01

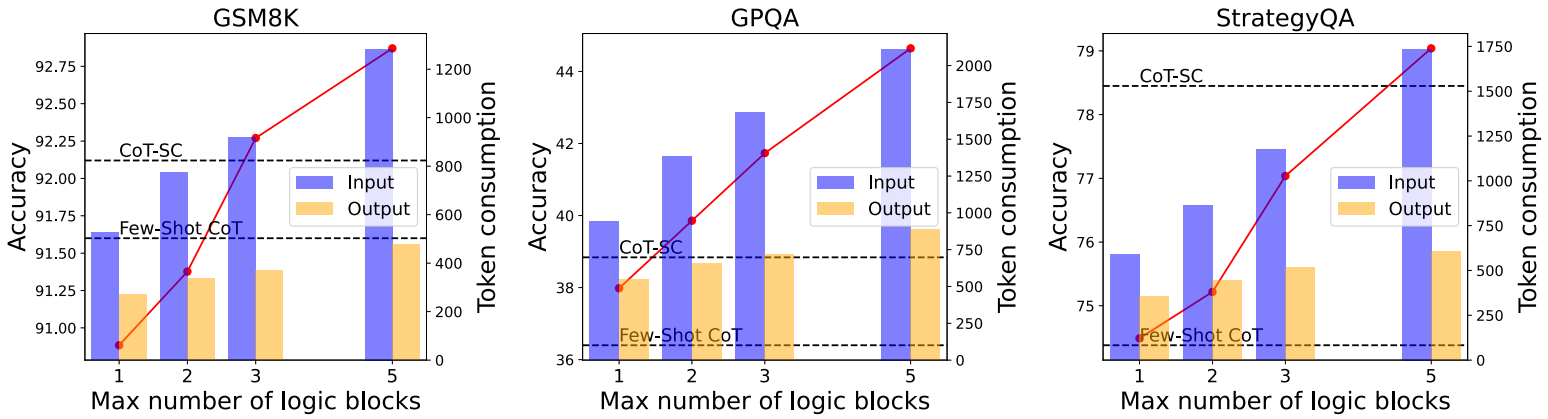

Figure 1: Test-time scaling with different number of logic blocks.

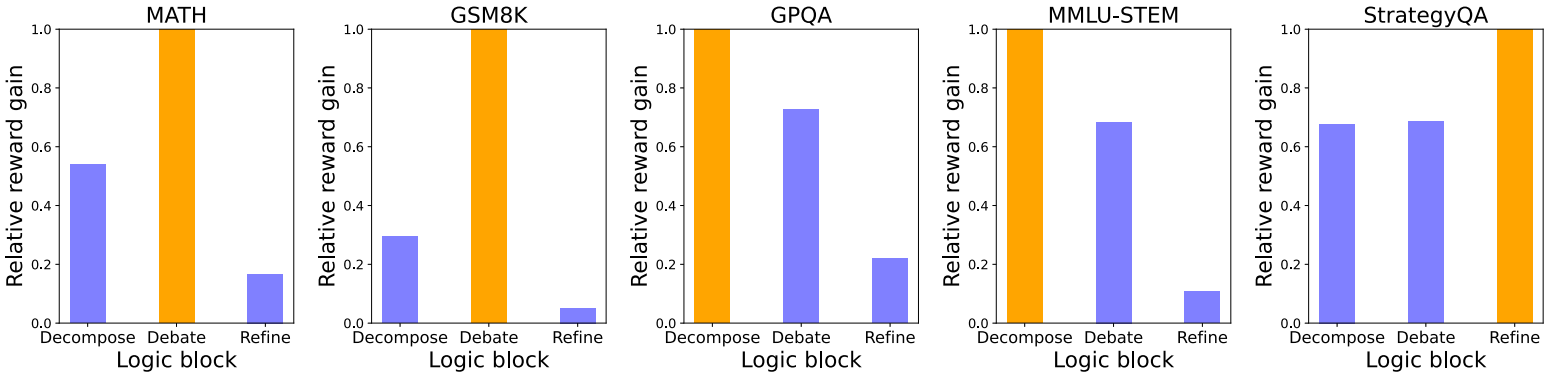

Figure 2: Contribution of each logic block in the reasoning sequence.