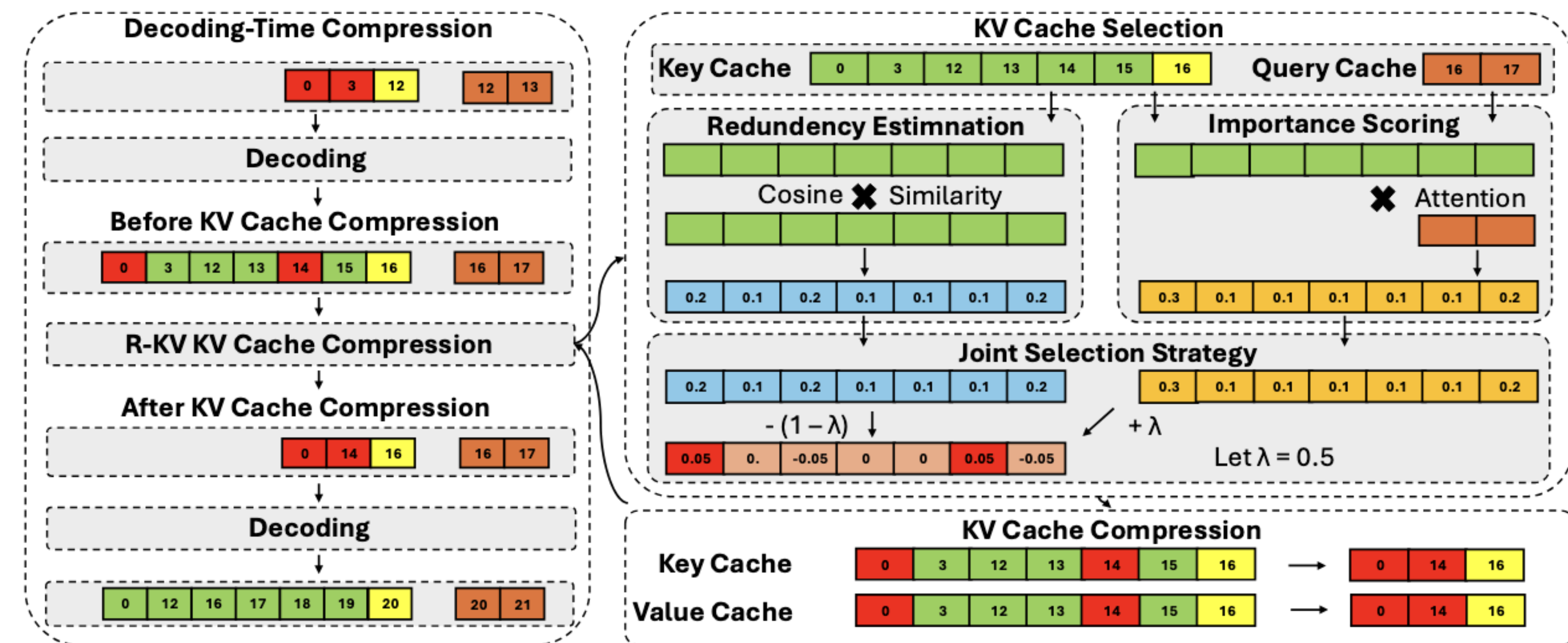
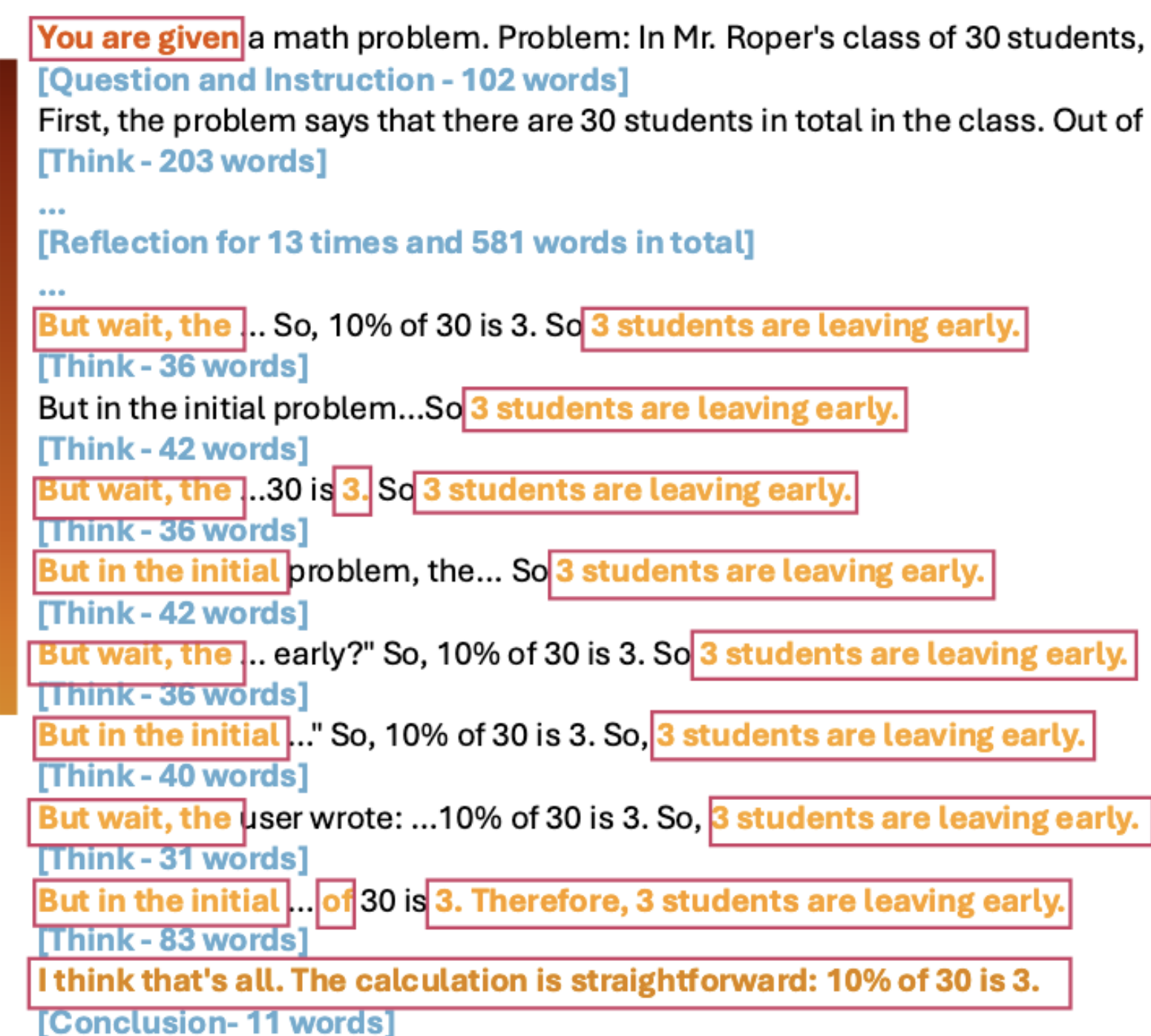


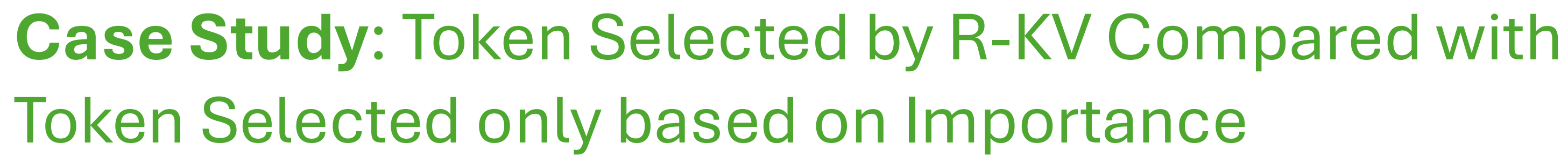


Overview: Decoding Time Compression (Left) and Redundancy and Importance Joint Selection (Right)



We want to make your Reasoning Model run fast!

Results: Match Full KV Baseline with 10% KV Cache Maintained



Efficiency Results: R-KV Achieves 8X Throughput Improvement with Lossless Performance



Gen. Length	Method	Budget	Mem. Saving (%)	Batch	Throughput (tok/s)	Tokens Gen.	Dec. Time (s)
8K	FullKV	–	–	1	75.44	8 094	107.30
		–	–	62 (max)	849.13	501 828	590.99
	R-KV	Fixed – 1024	87.50	1	80.46	8 094	100.60
		Fixed – 1024	87.50	402 (max)	3 251.52	3 253 788	1 000.70
		Fixed – 1536	81.25	287 (max)	2 525.75	6 546 972	919.72
		Ratio – 10% – 819	90.00	479 (max)	3 809.15	3 877 026	1 017.82
		Ratio – 34% – 2 785	66.00	167 (max)	1 608.01	1 351 698	840.61
		Ratio – 54% – 4 423	46.00	105 (max)	1 257.83	849 870	675.66
16K	FullKV	–	–	1	69.41	16 286	234.65
		–	–	30 (max)	347.03	488 580	1 407.89
	R-KV	Fixed – 1024	93.75	1	80.95	16 286	201.18
		Fixed – 1024	93.75	402 (max)	3 188.82	6 546 972	2 053.10
		Fixed – 1536	90.63	287 (max)	2 447.61	4 674 082	1 909.65
		Ratio – 10% – 1 638	90.00	271 (max)	2 300.28	4 413 506	1 918.68
		Ratio – 34% – 5 570	66.00	82 (max)	797.43	1 335 452	1 674.70
		Ratio – 54% – 8 847	46.00	46 (max)	584.77	749 156	1 281.12