

Configuration	Chunk size	Overlap	Avg. precision	Avg. relevance	P95 latency	Cost per 1K queries
---------------	------------	---------	-------------------	-------------------	-------------	------------------------

Config A	256 tokens	20%	0.52	0.74	1.8s	\$3.20
----------	------------	-----	------	------	------	--------

Config B	512 tokens	30%	0.61	0.81	2.1s	\$4.50
----------	------------	-----	------	------	------	--------

Config C	1024 tokens	40%	0.58	0.83	2.8s	\$6.10
----------	-------------	-----	------	------	------	--------