

Where do prefix cache hits come from? (N=1214 requests; block-aligned 512-tok blocks)



0%

25%

50%

75%

100%

■ intra-session reuse (79.2%)

■ cross-session reuse (0.8%)

■ first emission (reused later) (10.1%)

■ unique (never reused) (9.9%)