Table 4 MultiHop-RAG retrieval on 2,255 non-null queries (top-10).

Method	Any	All	Recall	ΔAll	95% CI	p_{Holm}
BM25	0.982	0.583	0.807	–	–	–
Dense BGE	0.979	0.503	0.767	-0.079	[-0.098, -0.060]	< 0.001
Qwen-HyDE	0.970	0.448	0.731	-0.134	[-0.155, -0.114]	< 0.001
Mistral-HyDE	0.968	0.434	0.718	-0.149	[-0.169, -0.128]	< 0.001
Cross-encoder	0.996	0.729	0.890	+0.146	[+0.127, +0.165]	< 0.001

Parent-document evidence IDs define support; 301 null queries are excluded from retrieval denominators. Cross-encoder reranks a frozen BM25 top-100 pool. All-support and EM are primary endpoints. Four observed contrasts versus BM25 use Holm correction with a fifth slot conservatively reserved at 1.0 for unavailable IRCoT; this is not an observed IRCoT p-value. The full registered family is incomplete. Intervals are 95% question-level bootstrap intervals (10,000 resamples); intervals alone do not determine significance.