Table 6: Comparison of query reformulation (QR) methods using different dense retrieval embedding models across three benchmark datasets. Results are averaged over five independent runs with the LLM (temperature = 1). Best-performing results are highlighted in bold. Statistically significant improvements over the strongest baseline (paired t-test, $p < 0.05$) are marked with an asterisk. **EQR** consistently and significantly outperforms other LLM-based QR methods across all datasets and embedding models.

	TravelDest				**TripAdvisor Hotel**				**Yelp Restaurant**			
	NDCG@10	NDCG@30	P@10	P@30	NDCG@10	NDCG@30	P@10	P@30	NDCG@10	NDCG@30	P@10	P@30
all-MiniLM-L6-v2												
No QR	0.565	0.506	0.548	0.476	0.259	0.306	0.205	0.154	0.473	0.437	0.455	0.381
GQR	0.588	0.531	0.573	0.492	0.305	0.341	0.244	0.170	0.521	0.472	0.500	0.415
GenQR	0.615	0.549	0.590	0.509	0.349	0.369	0.284	0.172	0.574	0.511	0.548	0.443
GenQREnsemble	0.607	0.533	0.567	0.501	0.333	0.376	0.269	0.169	0.561	0.503	0.538	0.425
Q2E	0.627	0.532	0.603	0.514	0.325	0.374	0.259	0.172	0.573	0.508	0.547	0.419
Query2Doc	0.673	0.554	0.640	0.510	0.288	0.343	0.218	0.164	0.424	0.385	0.402	0.347
EQR (Ours)	**0.720***	**0.620***	**0.677***	**0.563***	**0.369***	**0.408***	**0.295***	**0.185***	**0.616***	**0.536***	**0.583***	**0.466***
e5-small-v2												
No QR	0.578	0.522	0.564	0.491	0.269	0.314	0.216	0.152	0.577	0.505	0.545	0.427
GQR	0.632	0.557	0.614	0.516	0.293	0.347	0.226	0.168	0.602	0.534	0.567	0.443
GenQR	0.686	0.592	0.659	0.539	0.321	0.375	0.245	0.174	0.636	0.553	0.597	0.466
GenQREnsemble	0.661	0.590	0.641	0.516	0.337	0.368	0.250	0.159	0.629	0.545	0.570	0.471
Q2E	0.655	0.546	0.634	0.521	0.325	0.380	0.257	0.173	0.616	0.546	0.570	0.468
Query2Doc	0.692	0.579	0.647	0.514	0.282	0.340	0.224	0.166	0.525	0.470	0.500	0.416
EQR (Ours)	**0.721***	**0.613***	**0.690***	**0.558***	**0.358***	**0.398***	**0.273***	**0.178**	**0.656***	**0.571***	**0.618***	**0.491***