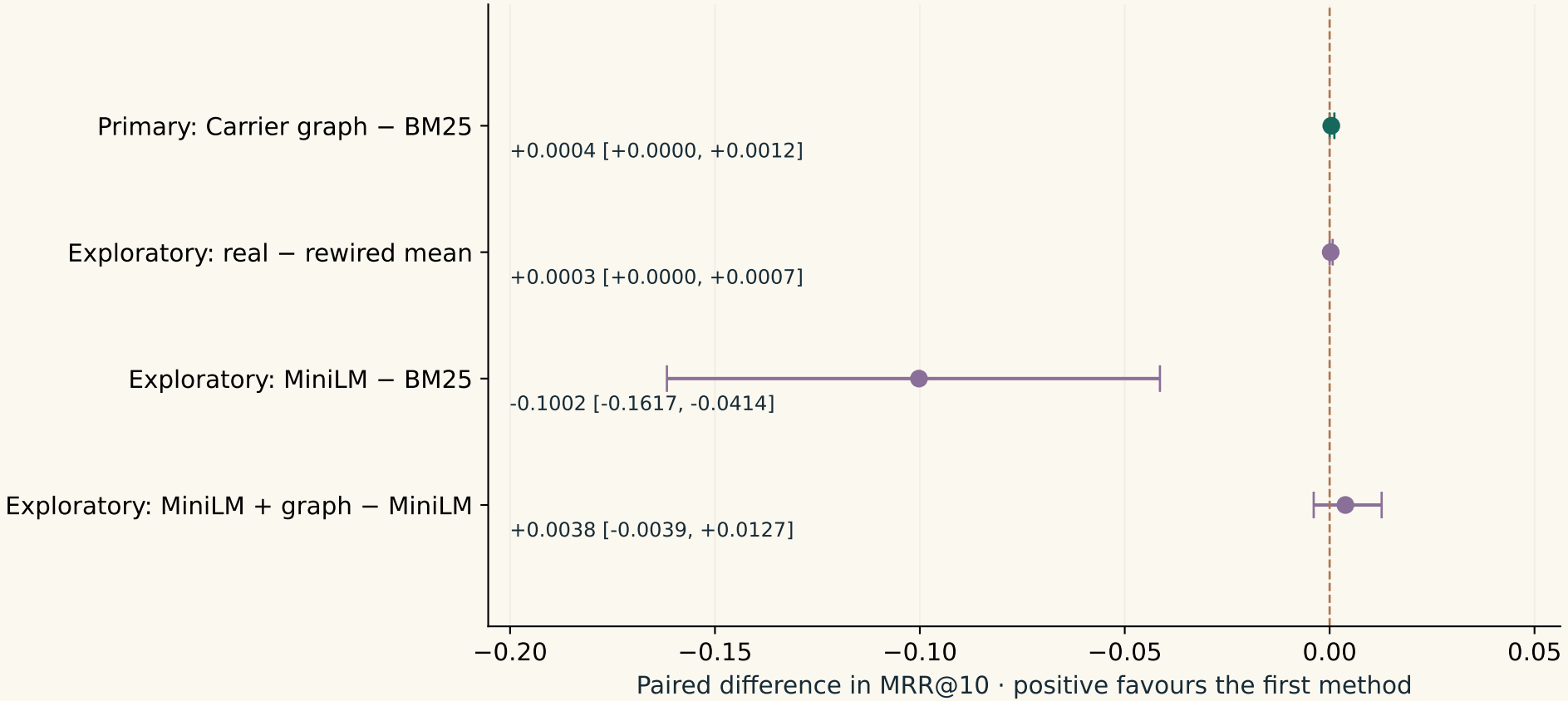
A small difference needs its uncertainty.

95% paired bootstrap intervals · 10,000 family-level resamples



Primary two-sided sign-flip p = 0.4944. Resampling unit: concept family, not query variant.
Secondary comparisons are exploratory. These intervals do not establish performance for real users.