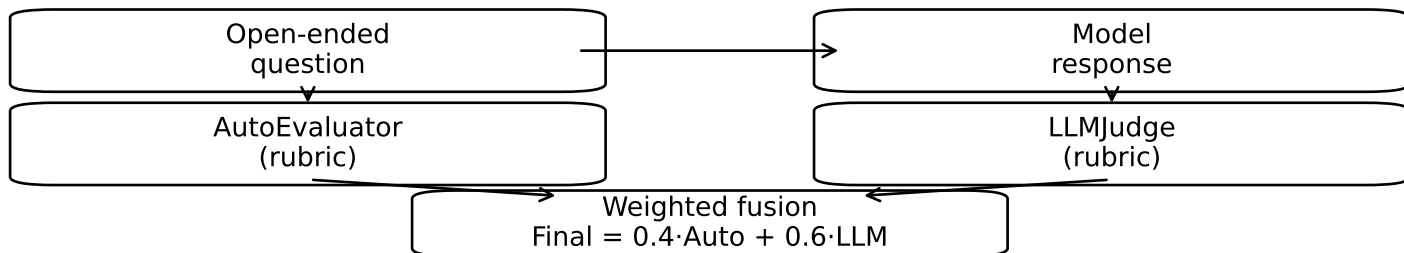
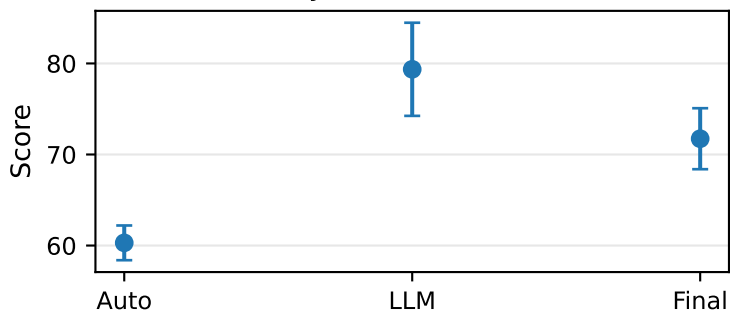


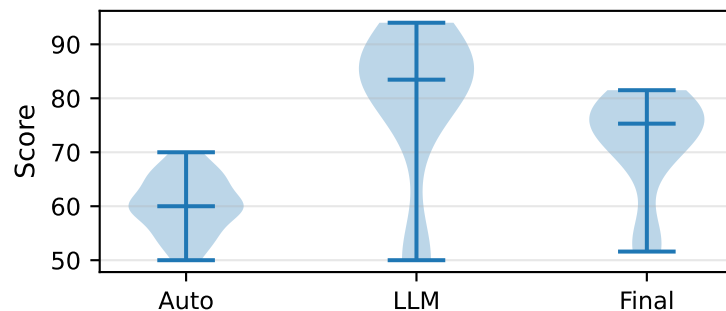
(a) Open-ended evaluation framework



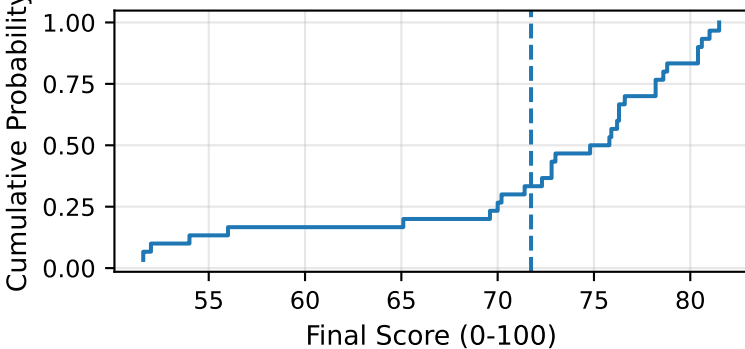
(b) Summary (mean \pm 95% CI, N=30)



(c) Score distributions (violin + median)



(d) ECDF of Final scores (mean=71.73)



(e) Score stability vs sample size

