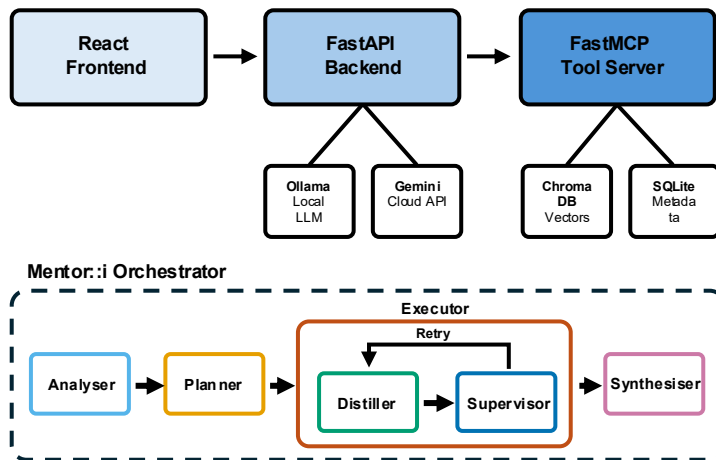


a



b

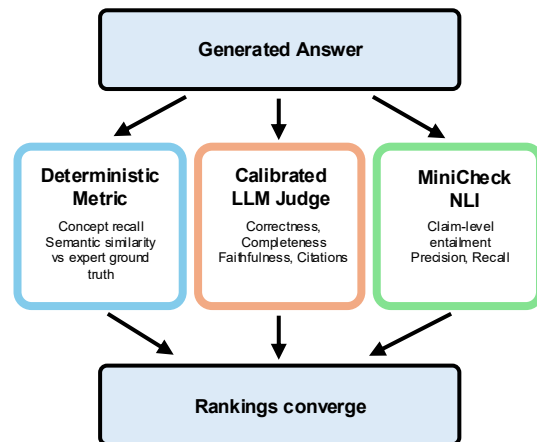
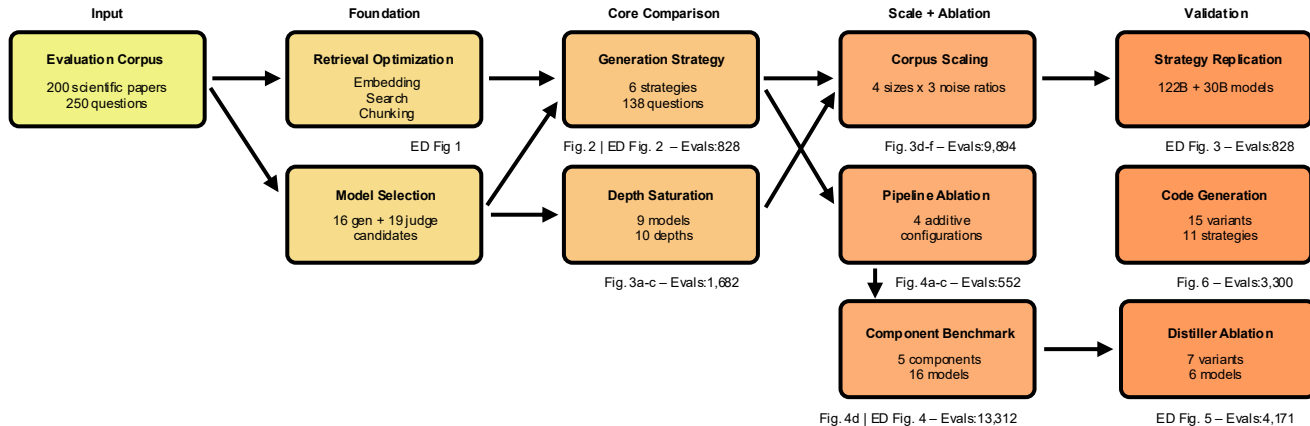


Fig. 5 | ED Fig. 6

c



**Total: 36,000+ evaluations across 200 papers, 250 questions, 10 experiments in 2 task domains (QA + code generation)**