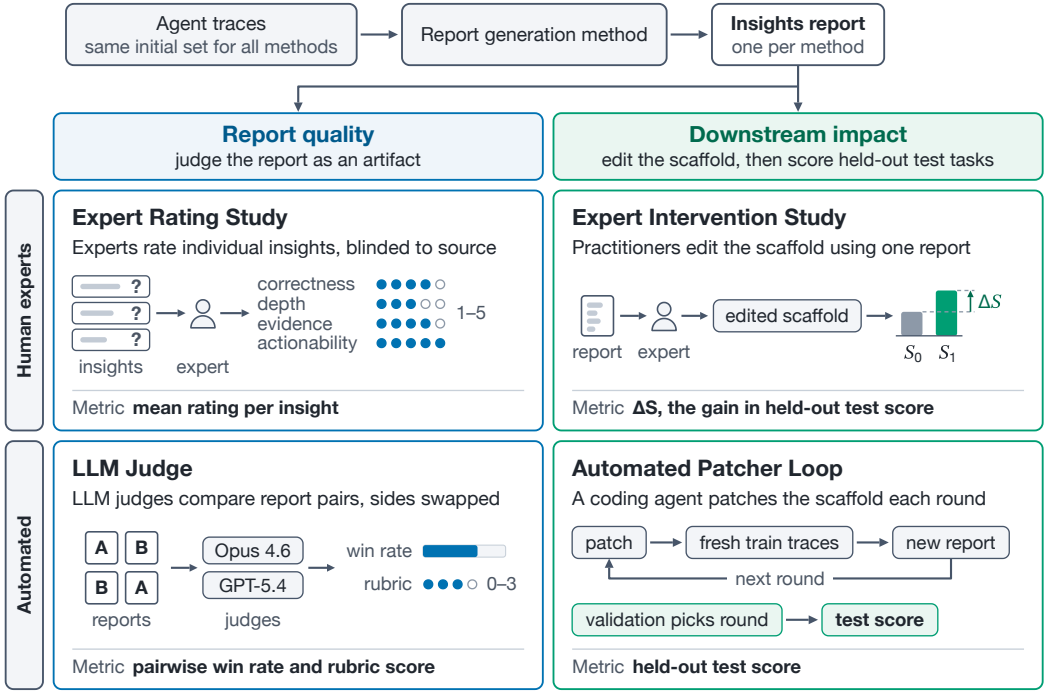

Agent traces
same initial set for all methods
Report generation method
Insights report
one per method
Report quality
judge the report as an artifact
Downstream impact
edit the scaffold, then score held-out test tasks
Human experts
Expert Rating Study
Experts rate individual insights, blinded to source
?
?
?
insights
expert
correctness
depth
evidence
actionability
1–5
Metric mean rating per insight
Expert Intervention Study
Practitioners edit the scaffold using one report
report
expert
edited scaffold
ΔS
S_0
S_1
Metric ΔS, the gain in held-out test score
Automated
LLM Judge
LLM judges compare report pairs, sides swapped
A
B
B
A
reports
Opus 4.6
GPT-5.4
judges
win rate
rubric
0–3
Metric pairwise win rate and rubric score
Automated Patcher Loop
A coding agent patches the scaffold each round
patch
fresh train traces
new report
next round
validation picks round
test score
Metric held-out test score