

Evaluation Four independent questions in human-centric reasoning.

Reference match Conditional Error



Opus 5
M Personalized reply
Preference mismatch

Raw 0% Verified 0%

Island street interview



10.0 s

Shares family history



20.0 s

Family story continues



28.0 s

Given preference favors less personal follow-up

Question How should EGO respond to the family story?
Given memory #4 “... appreciates minimal personal follow-up.”
Reference E: “Acknowledge the story warmly ... a wonderful family tradition.”
Prediction J: “Encourage him to continue by asking for more details ...”
Check The prediction adds personal follow-up despite the stated preference.



3.1 Pro
S Response strategy
Temporal mismatch

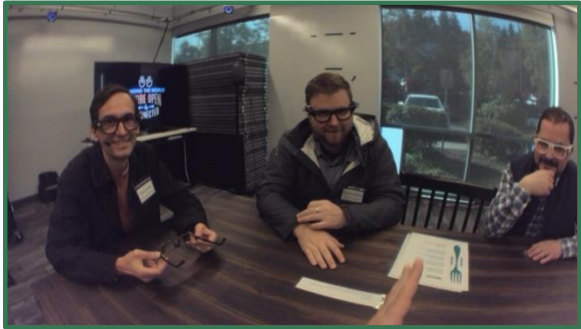
Raw 100% Verified 55.6%

Queried time window



14.5 s

Clarifies riddle constraints



18.5 s

Later cited question



23.5 s

Correct answer, evidence from a later window

Question Which strategy does EGO use during 14.1–18.65 s?
Reference H / Prediction H: ASK_CLARIFICATION
Evidence “Can he be on a mountain ...” (14.1–18.65 s)
Rationale (S) “Does the mile ... south direction?” (22.45–25.05 s)
Check Right answer, wrong supporting conversational window.



5.6 Sol
T Interaction timing
Timing disagreement

Raw 0% Verified 0%

Ongoing conversation



8.0 s

Gesture during speech



10.5 s

Later speech segment



16.5 s

Different anchors for the waiting decision

Question After the thumbs-up, when should EGO wait for the speaker?
Reference A: 10–12 s Prediction D: 16–18 s
Rationale (T) “The speaker completes the thought about being treated as a person around 16 seconds.”
Check GPT chooses a later speech boundary than the reference cue.



3.8-Max
N Request fulfillment
Goal substitution

Raw 0% Verified 0%

Cutting of redwoods



10.0 s

Why were they cut?



12.7 s

Wastewater pipe use



16.0 s

Fulfilled inquiry mistaken for an unmet goal

Question Was the tree-touching man's inquiry fulfilled?
Context / ask “They were cutting redwoods.” / “Why?”
Reply “...using them for pipes for wastewater ... before they had metal pipes.”
Reference F: MET_REQUEST Prediction E: UNMET_REQUEST
Rationale (S) “the desire to preserve them was unmet historically”