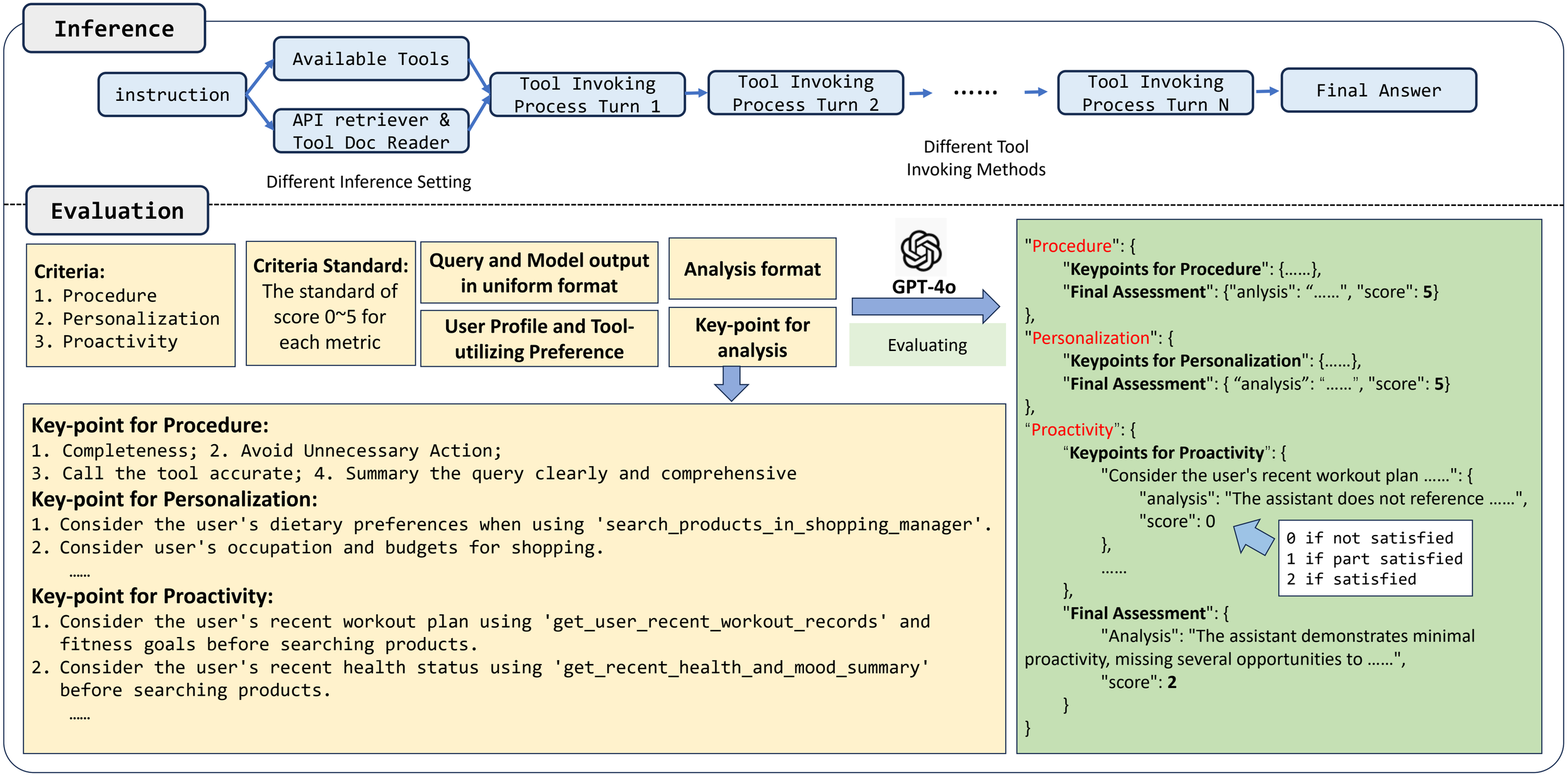
Inference
instruction
Available Tools
API retriever & Tool Doc Reader
Different Inference Setting
Tool Invoking Process Turn 1
Tool Invoking Process Turn 2
.....
Tool Invoking Process Turn N
Final Answer
Different Tool Invoking Methods
Evaluation
Criteria:
1. Procedure
2. Personalization
3. Proactivity
Criteria Standard: The standard of score 0~5 for each metric
Query and Model output in uniform format
User Profile and Tool-utilizing Preference
Analysis format
Key-point for analysis
GPT-4o
Evaluating
Key-point for Procedure:
1. Completeness; 2. Avoid Unnecessary Action;
3. Call the tool accurate; 4. Summary the query clearly and comprehensive
Key-point for Personalization:
1. Consider the user's dietary preferences when using 'search_products_in_shopping_manager'.
2. Consider user's occupation and budgets for shopping.
.....
Key-point for Proactivity:
1. Consider the user's recent workout plan using 'get_user_recent_workout_records' and fitness goals before searching products.
2. Consider the user's recent health status using 'get_recent_health_and_mood_summary' before searching products.
.....
"Procedure": {
"Keypoints for Procedure": {.....},
"Final Assessment": {"anlysis": ".....", "score": 5}
},
"Personalization": {
"Keypoints for Personalization": {.....},
"Final Assessment": { "analysis": ".....", "score": 5}
},
"Proactivity": {
"Keypoints for Proactivity": {
"Consider the user's recent workout plan .....": {
"analysis": "The assistant does not reference .....",
"score": 0
},
.....
},
"Final Assessment": {
"Analysis": "The assistant demonstrates minimal proactivity, missing several opportunities to .....",
"score": 2
}
}
0 if not satisfied
1 if part satisfied
2 if satisfied