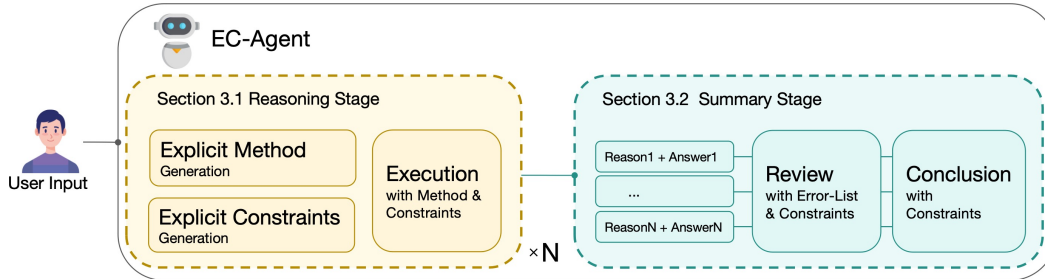


000 EXPLICIT-CONSTRAINED SINGLE AGENT FOR
 001 ENHANCED TASK-SOLVING IN LLMs
 002 APPENDIX
 003

004 APPENDIX A
 005



017 Figure 1: The pipeline of EC-Agent.

018
 019 **EC-Agent Prompt:**

020
 021 **Reasoning Stage Prompt:**

022 {Task}
 023 Step 1: Find a method.
 024 Step 2: List and analyze all the constraints.
 025 Step 3: Reason step by step, output the answer in the format: "Final Answer is:"
 026
 027

028
 029 **Summary Stage Prompt:**

030 **Review Module:**

031 {Potential Answer}
 032 {Task}
 033 Step 1: List common errors in this type of task.
 034 Step 2: Evaluate and Self-Reflection.
 035 Step 3: Reason and output modified result.
 036

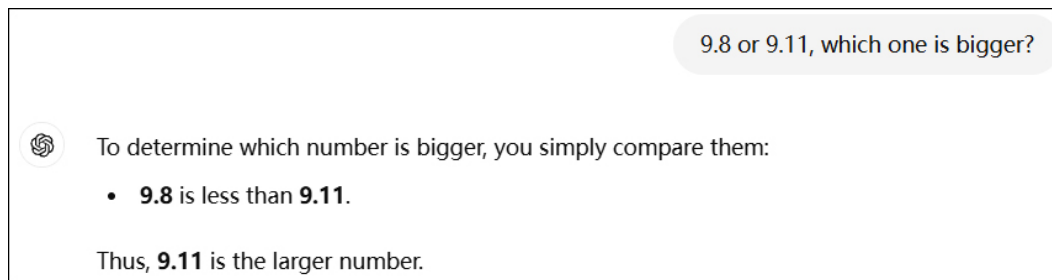
037 **Conclusion Module:**

038 {Potential Answer List}
 039 {Task}
 040 Step 1: Get result from potential answer, list and summary.
 041 Step 2: Make a major vote on all potential answers.
 042 Step 3: If all answers are incomplete, reason out a new one.
 043 Step 4: Output the final result.
 044
 045
 046
 047
 048
 049
 050
 051
 052
 053

054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107

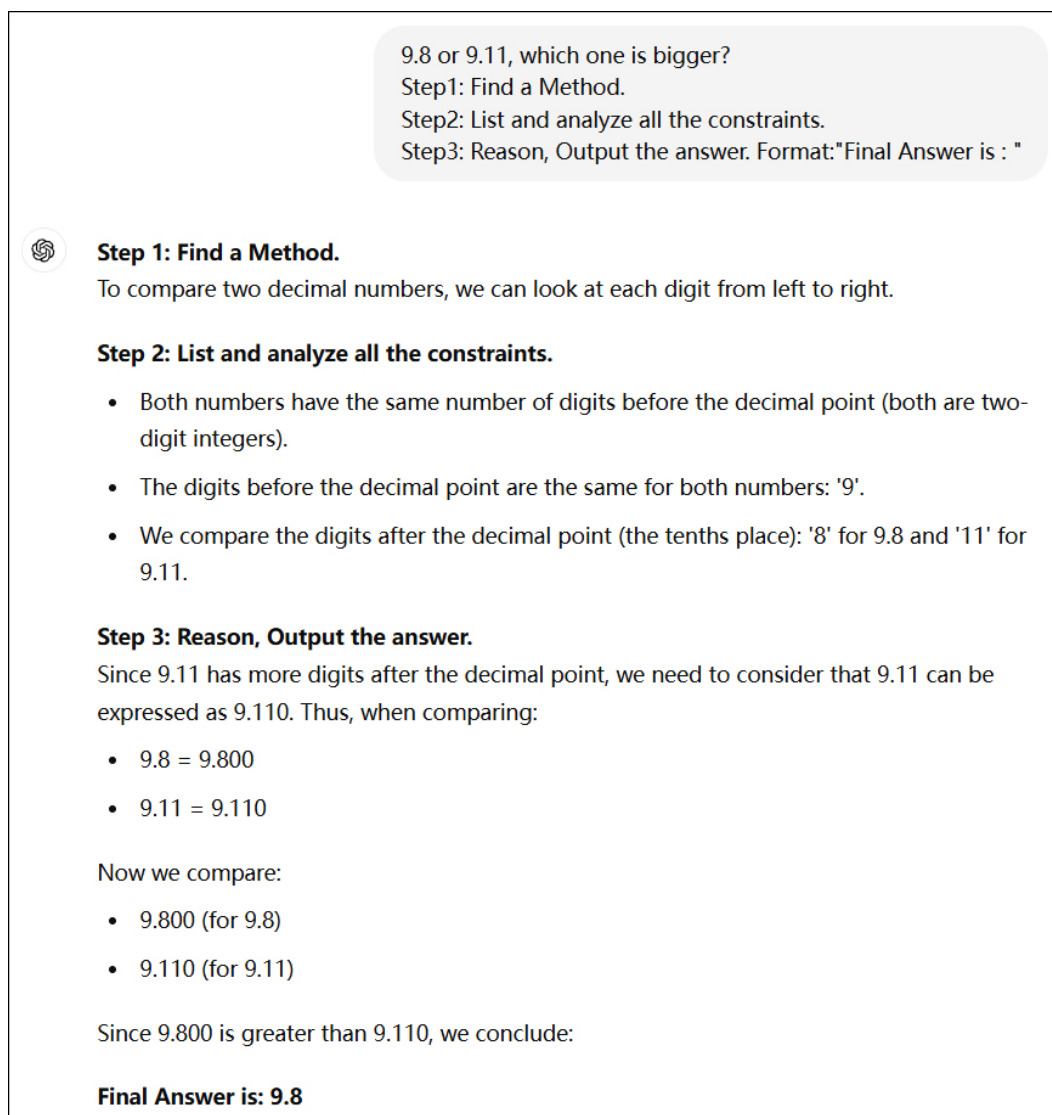
APPENDIX B

TASK 1: 9.8 OR 9.11, WHICH ONE IS BIGGER?



The screenshot shows a chat interface. At the top right, a grey bubble contains the prompt: "9.8 or 9.11, which one is bigger?". Below it, a response from GPT-4o is shown, starting with a circular icon. The response text is: "To determine which number is bigger, you simply compare them:" followed by a bulleted list: "• 9.8 is less than 9.11." and then "Thus, 9.11 is the larger number."

Figure 2: Reasoning by default prompt (GPT-4o).



The screenshot shows a chat interface. At the top right, a grey bubble contains the prompt: "9.8 or 9.11, which one is bigger?". Below it, a response from EC-Agent is shown, starting with a circular icon. The response text is: "Step 1: Find a Method. To compare two decimal numbers, we can look at each digit from left to right." followed by "Step 2: List and analyze all the constraints." and a bulleted list: "• Both numbers have the same number of digits before the decimal point (both are two-digit integers).", "• The digits before the decimal point are the same for both numbers: '9'.", and "• We compare the digits after the decimal point (the tenths place): '8' for 9.8 and '11' for 9.11." followed by "Step 3: Reason, Output the answer." and "Since 9.11 has more digits after the decimal point, we need to consider that 9.11 can be expressed as 9.110. Thus, when comparing:" followed by a bulleted list: "• 9.8 = 9.800" and "• 9.11 = 9.110". Then "Now we compare:" followed by a bulleted list: "• 9.800 (for 9.8)" and "• 9.110 (for 9.11)". Then "Since 9.800 is greater than 9.110, we conclude:" followed by "Final Answer is: 9.8".

Figure 3: Reasoning by EC-Agent (GPT-4o).

Correct reasoning method and conditional judgments can quickly guide the model to reach accurate conclusions.

108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161

TASK 2: EC-AGENT VS STANDARD PROMPT

Compared to the Standard Prompt, EC-Agent, through the control of Explicit Constraints and Explicit Method., enables the LLM to not only generate diverse execution methods and drive subsequent reasoning behaviors of the LLM, but also focuses on the requirements of the Task during the task execution process.

Task 2.1: Logic Grid Puzzle

Solve the task. Q. There are 4 houses in a row, numbered 1 on the left to 4 on the right. There is one person living in each house. The people in these houses have different characteristics.

- Each person has a different kind of car: one drives a convertible, one drives a minivan, one owns an SUV, and one drives a truck
- Each person ate something different for lunch: one had steak, one ate fried rice, one ate spaghetti, and one had pizza
- Each person has different flowers in their foyer: one has a vase of tulips, one has a rose bouquet, one has a bouquet of lilies, and one has a carnations arrangement
- Each person is eating a different kind of fruit: one is eating mangos, one is eating grapes, one is eating apples, and one is eating oranges
- Each person has a favorite color: one likes green, one likes white, one likes red, and one likes blue

Clues:

- The person who likes green lives in the first house
- The person who had pizza lives directly left of the person who likes blue
- The person who likes red lives somewhere to the right of the person who has a carnations arrangement
- The person who ate fried rice lives in the fourth house
- The person who is eating mangos lives directly left of the person who likes red
- The person who drives a truck does not live in the third house
- The person who ate spaghetti lives somewhere to the right of the person who drives a minivan
- The person who is eating mangos lives directly left of the person who had pizza
- The person who likes red lives somewhere to the left of the person who is eating grapes
- The person who had pizza and the person who has a rose bouquet live directly next to each other
- The person who likes blue and the person who owns an SUV live directly next to each other
- The person who had pizza and the person who is eating apples live directly next to each other

What is the number of the house where the person who had pizza lives?
Choose one correct room number from those "1,2,3,4."

| EC-Agent | Standard |
|---|---|
| <pre>## Solving the Puzzle **Step 1. Method** We'll use a combination of deduction and table-building to solve this puzzle. We'll create a table to track the information we know and use the clues to fill in the gaps. **Step 2. Analyzing Constraints** Here are the key constraints from the clues: - "Location": Houses numbered 1 to 4. - "Characteristics": Car type, lunch, flowers, fruit, favorite color - "Relationships": Direct left/right neighbors, specific orders. **Step 3. Reasoning** Let's start filling in our table based on the most direct clues: House Person Car Lunch Flowers Fruit Color --- --- --- --- --- --- --- 1 Green 2 3 4 4 Fried Rice 4 **Clue 1** Person in house 1 likes green. **Clue 4** Person in house 4 ate fried rice. **Clue 7** Truck driver lives left of lilies. We can continue filling in the table using these clues and their relationships. For example, clue 2 tells us the pizza eater is directly left of the blue-liker. Clues 9 and 13 help pinpoint the location of pizza and apples eater. ... (Omitting the reasoning text.) **Final Answer is: 2**</pre> | <pre>## Solving the Puzzle **Step 1. Method** We can use a systematic approach to solve this puzzle by creating a table and filling it in based on the clues. **Step 2. Analyzing Constraints** - "House Numbers." We have four houses numbered 1 to 4. - "Characteristics." Each house has a resident with unique characteristics (car, lunch, flowers, fruit, color). - "Clues." The clues provide direct and indirect relationships between these characteristics. **Step 3. Reasoning** Let's start by filling in the table based on the most direct clues: House Car Lunch Flowers Fruit Color --- --- --- --- --- --- 1 Green 2 3 4 Fried Rice 4 Now, we can use the remaining clues to deduce further information and fill in the table. For example: - Clue 2: Pizza is left of Blue. - Clue 6: Mangos are left of Red. - Clue 9: Mangos are left of Pizza, and so on. ... (Omitting the reasoning text.) **Final Answer is: 2**</pre> |
| <pre>**Step 1. Analyze Potential Answers** All three potential answers arrive at the same conclusion: **Answer 1** Final Answer is: 2 **Answer 2** Final Answer is: 2 **Answer 3** Final Answer is: 2 **Step 2. Major Vote** There's a unanimous agreement among the potential answers. **Step 3. No Need for Re-reasoning** Since all answers correctly identify the house number, we don't need to re-reason. **Step 4. Final Result** The person who had pizza lives in "house 2".</pre> | <p>Here's how to solve this logic puzzle:</p> <p>**1. Use the clues**</p> <p>**Clue 1:** The person who likes green lives in the first house.</p> <p>**Clue 4:** The person who ate fried rice lives in the fourth house.</p> <p>**2. Use Deduction**</p> <p>**Combine Clues:** Look for clues that link different pieces of information. For example, Clue 2 says the pizza eater is left of the blue-liker.</p> <p>**Elimination:** Eliminate possibilities for each house based on the clues.</p> <p>**3. Keep Iterating**</p> <p>Continue combining clues, and making deductions, until you can determine the house number where the person who had pizza lives.</p> |

Figure 4: Logic Grid Puzzle Task. The Blue Grid is a task query. The yellow Grid is reason result by EC-Agent, The green Grid is reason result by Standard Prompt.

In Figure4, the "EC-Agent" selected a method of reasoning that combines tables with conditional constraints, based on the requirements of the Task. In the Puzzle task, we compared and analyzed the failure reasons of tasks using the "Standard Prompt" and "EC-Agent". In "EC-Agent", the occurrence of early-termination in reasoning decreased by 60% compared to Standard Prompt (from 20 cases to 8 out of 100 tasks). We believe this is because, after applying the EM&EC instructions, the LLM's reasoning process became more focused.

162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215

Task 2.2: Codenames Collaborative

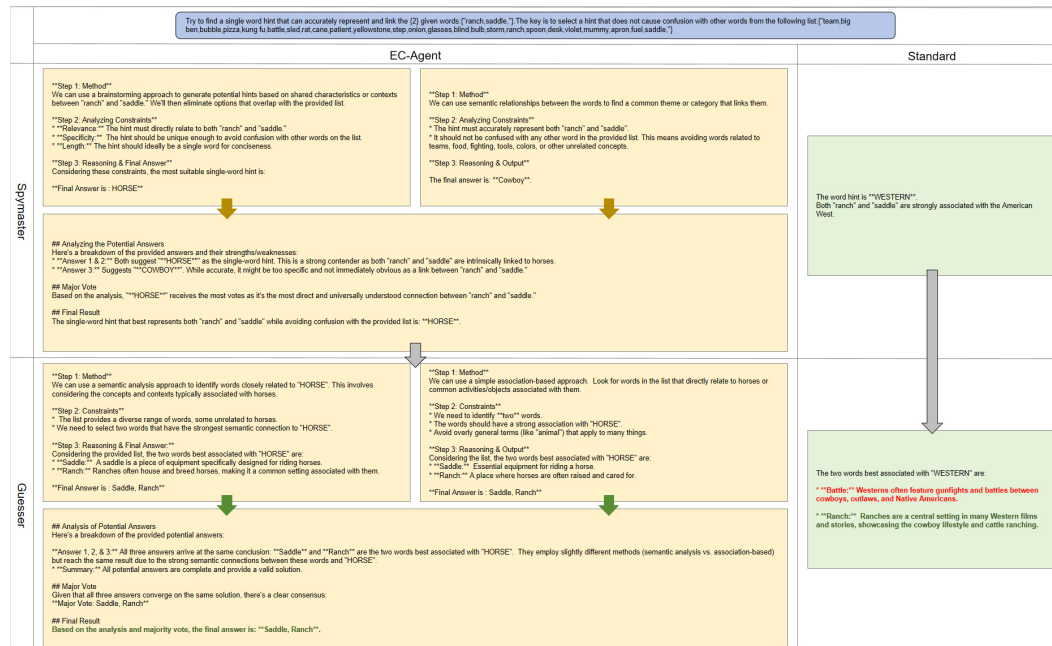


Figure 5: Codenames Collaborative. The Blue Grid is a task query. The yellow Grid is reason result by EC-Agent, The green Grid is reason result by Standard Prompt.

In Task 2.2, EC-Agent can specifically propose constraint conditions based on the Task requirements: "The hint should be unique enough to avoid confusion with other words on the list." And it follows this requirement during the reasoning and generation process: "Considering these constraints, the most suitable single-word hint is...". In Summary Stage, EC-Agent makes the correct choice for different "Potential Answers" by combining the requirements of the Task. In contrast, during the execution of the Standard Prompt, the LLM only considers the interrelationship between the two words "ranch" and "saddle" as required by the Task, ignoring the constraint relationships with other words, which leads to the failure of the final reasoning.

216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269

TASK 3: "EXPLICIT METHOD+EXPLICIT CONSTRAINTS" VS "PLAN"

| Try to find a single word hint that can accurately represent and link the [3] given words ("lab,needle,vet"). The key is to select a hint that does not cause confusion with other words from the following list. [speller slip: vet,needle,robin,trip,harm,fall,australia,spell,boost,jack,kangaroo,lab,web,lock,centaur,xtangle,table,dragon,pilot,ruler,battery,jam,cat.] | | |
|--|--|---|
| | EC-Agent | Standard |
| Spmyster | <p>""Step 1: Method"" We will analyze the relationships between the given words and identify common themes or actions associated with them. Then, we'll examine the provided list of words to find a single term that accurately reflects this theme without causing confusion with other entries.</p> <p>""Step 2: Constraints"" ""Accuracy"" The hint must directly relate to all three words: "lab," "needle," and "vet." ""Specificity"" The hint should be specific enough to avoid ambiguity and clearly point towards the connection between the words. ""Exclusion"" The hint cannot appear in the provided list of distractor words, ensuring no confusion or misinterpretation.</p> <p>""Step 3: Reasoning"" The common thread linking "lab," "needle," and "vet" is the field of "animal healthcare." Vets work in labs, often using needles for examinations and treatments.</p> <p>Considering the constraints, the most suitable single-word hint is:</p> <p>""Final Answer is: HEALTH""</p> | <p>""Step 1: Plan"" The plan is to identify a word that has a direct relationship with at least two of the given words ("lab," "needle," and "vet") without being too closely associated with any other words on the provided list.</p> <p>""Step 2: Reasoning"" 1. "Lab" is often associated with scientific experiments and medical procedures. 2. "Needle" is a tool used in both medical procedures and some scientific experiments. 3. "Vet" is a professional who performs medical procedures on animals, often using needles in labs or clinics.</p> <p>""Final Answer is: Examination""</p> |
| Guesser | <p>""Step 1: Method"" We can use a simple keyword-based association method. We'll look for words in the list that directly or indirectly relate to health, medicine, or well-being.</p> <p>""Step 2: Constraints"" ""Relevance"" The chosen words must have a clear connection to "health." ""Specificity"" We aim for words that are more specific to health than general concepts like "trip" or "table." ""Number"" We need to select only three words.</p> <p>""Step 3: Reasoning"" Considering the constraints, the following three words are best associated with "HEALTH":</p> <p>1. "Vet" - Veterinarians provide healthcare for animals. 2. "Needle" - Used in medical procedures like injections and blood draws. 3. "Lab" - Laboratories conduct tests related to health diagnosis and research.</p> <p>""Final Answer is: vet, needle, lab""</p> | <p>## Step 1: Plan To identify words best associated with "Examination", we'll look for words that relate to concepts like testing, assessment, or analysis. We'll consider synonyms, related actions, and tools used in examinations.</p> <p>## Step 2: Reasoning 1. "Lab": Laboratories are often associated with scientific examinations and experiments. 2. "Ruler": Rulers are used for measuring and assessing, which can be part of some examinations. 3. "Spell": Spelling tests are a common type of examination.</p> <p>Final Answer is: lab, ruler, spell</p> |

Figure 6: "Explicit Method+Explicit Constraints" vs "Plan". The Blue Grid is a task query. The yellow Grid is reason result by EC-Agent, The green Grid is reason result by "Plan" prompt.

EM+EC (left):

{ Task }

Step 1: Find a method.

Step 2: List and analyze all the constraints.

Step 3: Reason step by step, output the answer in the format: "Final Answer is:"

Plan+CoT (right):

{ Task }

Step 1: Make a plan.

Step 2: Reason step by step, output the answer in the format: "Final Answer is:"

We tested the effect of replacing the prompt keyword "EM&EC" with "Plan," and observed a decline in evaluation results. Our analysis suggests that this is because using the "EM&EC" combination allows the LLM to focus more on the core objectives of the task while reducing reliance on the correctness of the plan. The table below records the test results using different prompts.

270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323

APPENDIX C

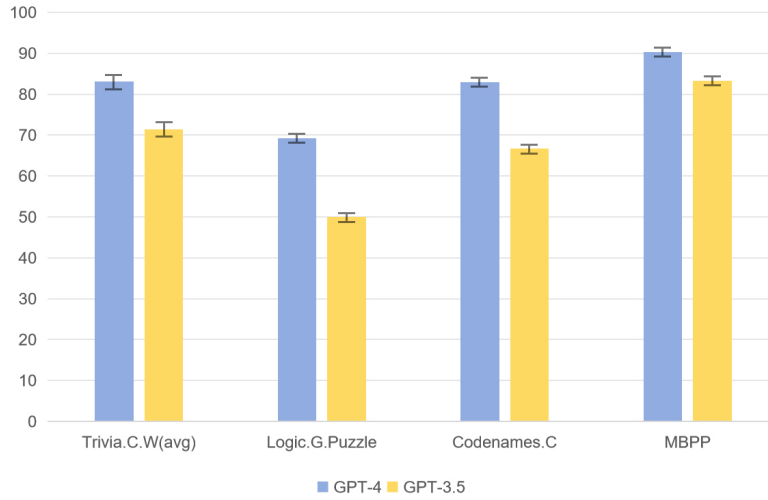


Figure 7: The scores and score variances of different tasks performed by the EC-Agent on OpenAI GPT series models.

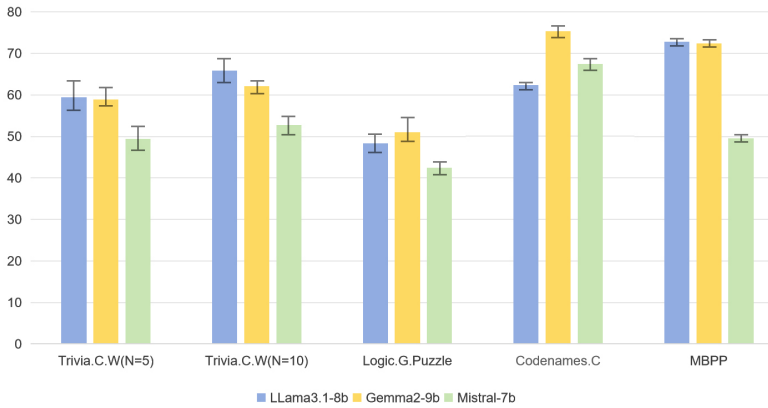


Figure 8: The scores and score variances of different tasks performed by the EC-Agent on multiple open-source models.

Combining numerical analysis, within the EC-Agent framework, some open-source models have reached or slightly surpassed the OpenAI GPT-3.5-0613 model on certain reasoning tasks (e.g., Logic G. Puzzle, Codename C.). However, there remains a significant gap to GPT-4-0613 overall. Additionally, different models also show variations in reasoning stability.