medians of two independent samples. Compared to the Analysis of Variance (ANOVA), it does not require the data to be normally distributed, making it suitable for small samples or data with unknown distribution.

The p-value of the test is 0.0024, which is below the commonly accepted significance level of 0.05. Therefore, we have reason to reject the null hypothesis, indicating that the "3R-Lp+PoT" strategy leads to a significant performance improvement.

A.5 Additional Error Types

Logic. In Table 4.2, we present an example of the logic error made by $\boldsymbol{R}^3$. We notice that LLMs may solve the problems using a more complicated logic, which is prone to mistakes. For instance, in Table 4.2, instead of directly counting the owners who do not own dogs, the LLMs try to substract the number of dog owners from the total number of owners. This ignores the possibility that some owners may have never had any dogs before. This addresses an issue with the multi-agent system that if the system comes up with a complicated initial SQL query, the following discussion process may try to polish the complicated SQL query instead of switching to an easier solution. In cases like Table 4.2, there is no way to reach a perfect SQL query with the substraction logic.

Inaccuracy. We observe that the LLMs may incorporate more information than what is asked by the end user. For instance, in Table 4.4, the user does not ask for the name of the dogs but the LLMs present such information along with the asked arriving date. We hypothesize that since such extra information can potentially be helpful to the end user, LLMs may be biased towards including it.

A.6 Spider Error Cases

Error Type	Question, Gold & Prediction	Reason
DB Value	**Q:** Find the last name of the students who currently live in the state of North Carolina but have not registered in any degree program. **Gold:** `SELECT ... WHERE T2.state_province_county = 'NorthCarolina' EXCEPT ...` **Pred:** `SELECT ... WHERE T2.state_province_county = 'North Carolina' EXCEPT ...`	The filtering condition in the question does not match the database value, string "NorthCalifornia" in database do not have a space in between.
Gold Error	**Q:** What are the first names of all players, and their average rankings? **Gold:** `SELECT avg(ranking), T1.first_name FROM players AS T1 JOIN rankings AS T2 ON T1.player_id = T2.player_id GROUP BY T1.first_name` **Pred:** `SELECT avg(ranking), T1.first_name FROM players AS T1 JOIN rankings AS T2 ON T1.player_id = T2.player_id GROUP BY T1.player_id`	The individuals in the table can be uniquely determined by column player_id not first_name, when `GROUP BY`.
Gold Error	**Q:** Find the id and cell phone of the professionals who operate two or more types of treatments. **Gold:** `SELECT T1.professional_id, T1.cell_number FROM Professionals AS T1 JOIN Treatments AS T2 ON T1.professional_id = T2.professional_id GROUP BY T1.professional_id HAVING count(*) >= 2` **Pred:** `SELECT T1.professional_id, T1.cell_number FROM Professionals AS T1 JOIN Treatments AS T2 ON T1.professional_id = T2.professional_id GROUP BY T1.professional_id HAVING COUNT( DISTINCT T2.treatment_type_code) >= 2`	The gold only finds professionals who have two or more records in the treatment table does not ensure that the records are for different types of treatments
Ambiguity	**Q:** What are the names and ids of all makers with more than 3 models? **Gold:** `SELECT T1.FullName, T1.Id FROM CAR_MAKERS AS T1 JOIN MODEL_LIST AS T2 ON T1.Id = T2.Maker GROUP BY T1.Id HAVING count(*) > 3;` **Pred:** `SELECT T1.Maker, T1.Id FROM CAR_MAKERS AS T1 JOIN MODEL_LIST AS T2 ON T1.Id = T2.Maker GROUP BY T1.Id HAVING count(*) > 3;`	Both column "Maker" and column "FullName" can answer the question about the "names of makers" in the query.
Imprecise	**Q:** What are the arriving date and the departing date of the dogs who have gone through a treatment? **Gold:** `SELECT DISTINCT T1.date_arrived, T1.date_departed FROM Dogs AS T1 JOIN Treatments AS T2 ON T1.dog_id = T2.dog_id` **Pred:** `SELECT DISTINCT T1.date_arrived, T1.date_departed, T1.Name FROM Dogs AS T1 JOIN Treatments AS T2 ON T1.dog_id = T2.dog_id`	The question do not require listing the specific names of the students, but only ask to list the students' arrival and departure dates. This falls under information redundancy.

Table 6

A.7 BIRD Error Cases

Error Type	Question, Gold & Prediction	Reason
DB Value	**Q:** How many cards with unknown power that can't be found in foil is in duel deck A? **Gold:** `SELECT SUM(CASE WHEN power LIKE '%*%' OR power IS NULL THEN 1 ELSE 0 END) FROM cards WHERE hasFoil = 0 AND duelDeck = 'a'` **Pred:** `SELECT COUNT(*) FROM cards WHERE (power IS NULL OR power = '*') AND hasFoil = 0 AND duelDeck = 'a'`	Values in database cannot exact match with the question and evidence.
Gold Error	**Q:** How many artists have designed a card with a black border color and is available in both "arena" and "mtgo" printing type? **Gold:** `SELECT COUNT(CASE WHEN availability LIKE '%arena,mtgo%' THEN 1 ELSE NULL END) FROM cards` **Pred:** `SELECT COUNT(DISTINCT artist) FROM cards WHERE borderColor = 'black' AND availability LIKE '%arena,mtgo%'`	The gold missed one filtering condition.
Ambiguity	**Q:** Among black card borders, which card has full artwork? **Gold:** `SELECT id FROM cards WHERE borderColor = 'black' AND isFullArt = 1` **Pred:** `SELECT name FROM cards WHERE borderColor = 'black' AND isFullArt = 1;`	Both column "name" and column "id" can answer the question about the "the card has full artwork" in the query.
Logic	**Q:** Which cities have the top 5 lowest enrollment number for students in grades 1 through 12? **Gold:** `SELECT T2.City FROM frpm AS T1 INNER JOIN schools AS T2 ON T1.CDSCode = T2.CDSCode GROUP BY T2.City ORDER BY SUM(T1.`Enrollment (K-12)`) ASC LIMIT 5` **Pred:** `SELECT T2.City, T1."Enrollment (K-12)" FROM frpm AS T1 JOIN schools AS T2 ON T1.CDSCode = T2.CDSCode ORDER BY T1."Enrollment (K-12)" ASC LIMIT 5;`	The Pred query is incorrect because it lacks grouping and aggregation. It directly selects cities and their corresponding "Enrollment (K-12)" without summing up the enrollment numbers across all schools within each city. As a result, it returns the cities of the five schools with the lowest enrollment numbers, which does not accurately answer the question seeking the top 5 cities with the lowest total enrollment for students in grades 1 through 12.

Table 7