Error Types	Question, Gold & Prediction	Explanation
Gold Error (30.5%)	**Q:** What are the Asian countries which have a population **larger than that of any country in Africa**? **Gold:** ❌ … `AND population > (SELECT min(population) FROM country WHERE Continent = "Africa")` **Pred:** ✅ … `AND population > (SELECT max(population) FROM country WHERE Continent = "Africa")`	Judged as incorrect because of the incorrect gold SQL query.
Logic (29.8%)	**Q:** How many owners **temporarily do not** have any dogs? **Gold:** ✅ `SELECT count(*) FROM Owners WHERE owner_id NOT IN (SELECT owner_id FROM Dogs)` **Pred:** ❌ `SELECT (SELECT COUNT(DISTINCT owner_id) FROM Owners) - (SELECT COUNT(DISTINCT owner_id) FROM Dogs WHERE date_departed IS NULL)`	The predicted SQL query wrongly assumes that all owners have had dogs.
Ambiguity (13.2%)	**Q:** What are the **names** of all makers with more than 3 models? **Gold:** ✅ `SELECT T1.FullName ... HAVING count(*) > 3;` **Pred:** ✅ `SELECT T1.Maker ... HAVING count(*) > 3;`	Both `FullName` and `Maker` columns hold the information for "names".
Inaccuracy (11.3%)	**Q:** What are the **arriving date** of the dogs who have gone through a treatment? **Gold:** ✅ `SELECT T1.date_arrived, FROM ...` **Pred:** ❌ `SELECT T1.date_arrived, T1.Name FROM ...`	The selected `Name` is not asked by the question.
DB Value (10.6%)	**Q:** Which city and country is the **Alton** airport at? **Gold:** ✅ `SELECT ... WHERE AirportName = "Alton" ;` **Pred:** ❌ `SELECT ... WHERE AirportName LIKE "%Alton%" ;`	Our framework notices there is a space for `Alton` in the DB, therefore employing a fuzzy match.
Others (4.6%)		

Table 4: Error Analysis of ***R***^{3} on Spider-Dev. We make the part in the question red when it is either annotated incorrectly in the gold SQL query (Gold) or predicted incorrectly in the predicted SQL query (Pred).

gold SQL queries, we still adopt the original set to calculate the performance of our system to ensure a fair comparison.

Gold Error. We notice that though the annotation quality of Spider is good, there are still cases where the gold SQL queries are not correct. Specifically, among the 151 examples, 30.5% are due to incorrect gold SQL queries (4.5% of all the examples in Spider-Dev). To facilitate future research, we catalog the instances with incorrect gold SQL, correct the errors, and share the details[3].

Ambiguity. We observe that there are a few questions involving ambiguities, a phenomenon spotted on a wide range of NLP tasks (Plank, 2022; Deng et al., 2023). In Table 4.3, both `FullName` and `Maker` columns hold the information for the "name of makers", except that `FullName` holds the full names while `Maker` holds the name abbreviations. Therefore, both the gold and predicted SQL queries should be considered correct if there is no further clarifications. Such ambiguous requests may be common in real-world applications as the lay users may not be familiar with the database schema. This requires future research on interactive Text-to-SQL systems that can understand and deal with such ambiguities in user questions.

[3] visible-after-review.com

Dirty Database Value. We observe that due to the Database (DB) setup for Spider, certain DB values may deviate from what is asked in the question. For instance, in Table 4.5, ***R***^{3} notices a space for `Alton` in DB, therefore employing a fuzzy match. But this deviates the SQL query's execution results from the gold SQL query's results.

Explanations of "Logic" and "Inaccuracy" errors can be found in Appendix A.5. Our findings indicate that the existing evaluation protocols for Text-to-SQL generation may not authentically capture the capabilities of these sophisticated systems. Therefore, we advocate for a reassessment and enhancement of Text-to-SQL evaluation methods. We provide further error analysis of ***R***^{3} on Bird in Appendix A.7.

4 Conclusion

R^3 significantly enhance the performances of LLMs on the Text-to-SQL task. We conduct a comprehensive error analysis and identify persistent issues with the current Text-to-SQL evaluation. This underscores the necessity for our community to develop a refined evaluation protocol that more effectively captures nuances in SQL generation and accurately reflects model performance.

Limitations

Due to the scope of the study, we only test a limited number of LLMs. The performance gap between 1R-Lp and 3R-Lp demonstrates that the number of reviewers is a worthwhile topic of research. However, this work does not delve into this much.

Ethical Statements

In this paper, we propose strategies to improve the SQL generation capabilities of LLMs. To the best of our knowledge, we do not expect our system would have negative impacts on the society.

References

AI@Meta. 2024. Llama 3 model card.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

Naihao Deng, Yulong Chen, and Yue Zhang. 2022. Recent advances in text-to-SQL: A survey of what we have and what we expect. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2166–2187, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

Xuemei Dong, Chao Zhang, Yuhang Ge, Yuren Mao, Yunjun Gao, Jinshu Lin, Dongfang Lou, et al. 2023. C3: Zero-shot text-to-sql with chatgpt. *arXiv preprint arXiv:2307.07306*.

Dawei Gao, Haibin Wang, Yaliang Li, Xiuyu Sun, Yichen Qian, Bolin Ding, and Jingren Zhou. 2023. Text-to-sql empowered by large language models: A benchmark evaluation. *arXiv preprint arXiv:2308.15363*.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

Zijin Hong, Zheng Yuan, Qinggang Zhang, Hao Chen, Junnan Dong, Feiran Huang, and Xiao Huang. 2024. Next-generation database interfaces: A survey of llm-based text-to-sql. *arXiv preprint arXiv:2406.08426*.

George Katsogiannis-Meimarakis and Georgia Koutrika. 2023. A survey on deep learning approaches for text-to-sql. *The VLDB Journal*, 32(4):905–936.

Jinyang Li, Binyuan Hui, Ge Qu, Binhua Li, Jiaxi Yang, Bowen Li, Bailin Wang, Bowen Qin, Rongyu Cao, Ruiying Geng, et al. 2023. Can llm already serve as a database interface. *A big bench for large-scale database grounded text-to-sqls. CoRR abs/2305.03111*.

Zhishuai Li, Xiang Wang, Jingjing Zhao, Sun Yang, Guoqing Du, Xiaoru Hu, Bin Zhang, Yuxiao Ye, Ziyue Li, Rui Zhao, et al. 2024. Pet-sql: A prompt-enhanced two-stage text-to-sql framework with cross-consistency. *arXiv preprint arXiv:2403.09732*.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mohammadreza Pourreza and Davood Rafiei. 2023. Din-sql: Decomposed in-context learning of text-to-sql with self-correction. *arXiv preprint arXiv:2304.11015*.

Matthew Renze and Erhan Guven. 2024. The effect of sampling temperature on problem solving in large language models. *arXiv preprint arXiv:2402.05201*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Bing Wang, Changyu Ren, Jian Yang, Xinnian Liang, Jiaqi Bai, Qian-Wen Zhang, Zhao Yan, and Zhoujun Li. 2023. Mac-sql: Multi-agent collaboration for text-to-sql. *arXiv preprint arXiv:2312.11242*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. 2023. Language agents with reinforcement learning for strategic play in the werewolf game. *arXiv preprint arXiv:2310.18940*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *arXiv preprint arXiv:1809.08887*.

Ruiqi Zhong, Tao Yu, and Dan Klein. 2020. Semantic evaluation for text-to-sql with distilled test suites. *arXiv preprint arXiv:2010.02840*.

A Appendix

A.1 Dataset Descriptions

	Spider-Dev	Spider-Test	Bird-Dev
	(Yu et al., 2018)		(Li et al., 2023)
#QA	1,034	2147	1,534
#Domain	138	-	37
#DB	200	206	95
DB Size	879.5 MB	906.5 MB	1.76 GB

Table 5: Statistics of two Text-to-SQL benchmarks we use in our experiments. "#QA", "#Domain" and "#DB" refer to the number of samples, domains and databases, respectively.

A.2 Baseline

Experiments in this work was based on LLMs including GPT-3.5-Turbo, GPT-4 (OpenAI, 2023) and Llama-3 (AI@Meta, 2024). As for the compared methods, the raw performance for GPT-3.5 ("-") was evaluated by Li et al. (2023); C3 employs schema linking filtering (Dong et al., 2023); DAIL selects few-shot demonstrations based on their skeleton similarities (Gao et al., 2023), and "SC" represents Self-Consistency (Wang et al., 2022); PET uses cross-consistency (Li et al., 2024); DIN decomposes the text-to-SQL task into smaller subtasks (Pourreza and Rafiei, 2023); MAC, as previously mentioned, is the first to apply a Multi-Agent system to Text-to-SQL tasks (Wang et al., 2023).

A.3 Effects of k in k-shot.

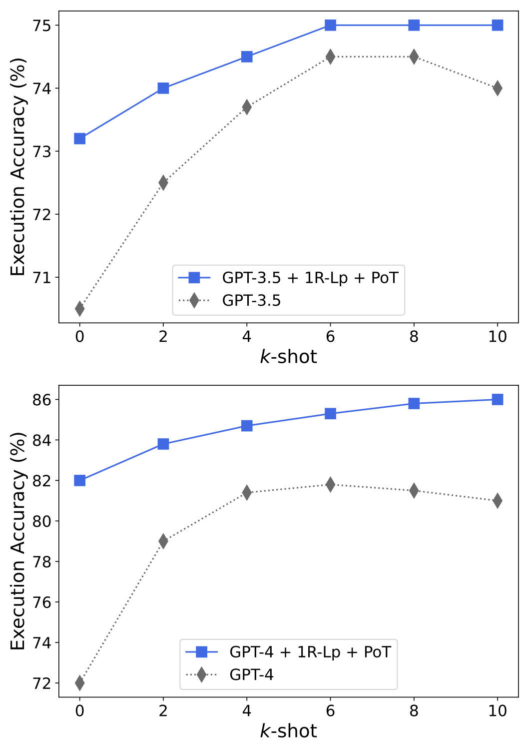

Figure 2: k-shot Sensitivity Analysis.

We test various k values on 200 random samples from Spider-Dev. As shown in Figure 2, compared to CoT, the performance of the $\boldsymbol{R}^3$ system remains relatively stable regardless of the number of examples, which corroborates our previous findings from the 0-shot experiments with Llama-3.

A.4 Significance Test

We divided the generated SQL by several strategies in Table 3 into 10 equal parts and calculated the execution accuracy for each. To test whether our strategy can indeed improve execution accuracy, we conduct a significance test between the "CoT" and "3R-Lp+PoT" strategies. The null hypothesis of the test is that the median execution accuracy obtained by the two strategies is the same. The Mann-Whitney U Test (Mann and Whitney, 1947) is a non-parametric statistical method used to compare whether there is a significant difference in the