

## 1 A Technical Appendices and Supplementary Material

### 2 A.1 Fine-Grained Analysis

3 We present a more fine-grained analysis by (1) investigating the performance changes over the SQLs  
 4 with varied length; and (2) counting the translation errors raised by different LLMs. As shown in  
 5 Figure 1, we observe that all the LLMs encounter performance regression when the SQLs evolve  
 6 to be more lengthy. Specifically, all the LLMs exhibit an average performance degradation when  
 7 the number of tokens involved in the SQL increases from 0 – 402 to 1214 – 2182. Moreover, these  
 8 LLMs showcase diverse translation errors across different SQLs (e.g., BAD\_ARGUMENTS indicates  
 9 that LLMs utilize incorrect arguments like the mismatched data types in the translated SQLs). The  
 10 results can be attributed to two aspects: (1) longer queries typically involve more operations to be  
 11 resolved, thus increasing the translation difficulty; (2) lengthy queries increase the risk of triggering  
 12 the limitation of LLMs, including the hallucination and lost-in-the-middle problem. Therefore, it  
 13 calls for techniques to enable LLMs to perform accurate translation over lengthy SQLs.

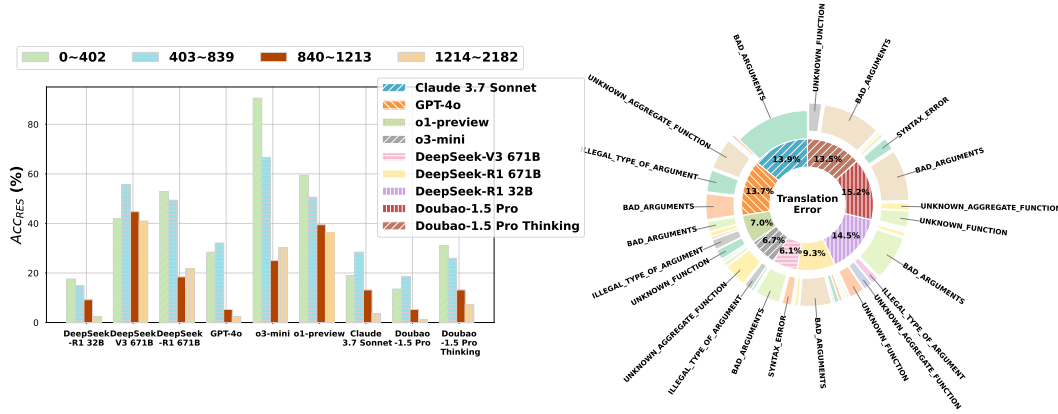


Figure 1: **Left:** Distribution of Translation Accuracy (%) over SQL Length. **Right:** Error Distribution over LLM Translation Results (The error information is collected based on the database feedback).

### 14 A.2 Typical Translation Type

15 **PARROT** categorizes cross-system SQL translation challenges into several common types based on  
 16 structural, lexical, and functional differences across database systems. The typical translation types  
 17 currently investigated in **PARROT** are listed in Table 1.

Table 1: Typical Translation Types in **PARROT**.

Translation	Description
<b>Syntax Rule</b>	Differences in syntactic structure requirements across databases.
<b>Keyword</b>	Naming differences for reserved words or functional keywords.
<b>Data Type</b>	Naming or precision differences for equivalent logical data types.
<b>Operator &amp; Built-in Function</b>	Name/behavior differences for operators or built-in functions.
<b>Stored Procedure</b>	Differences in definition and invocation syntax.
<b>UDF</b>	Differences in creation and usage of user-defined functions.
<b>Other</b>	Miscellaneous special differences (e.g., variable prefixes, comment symbols).

### 18 A.3 Dataset Details

19 We present the details of the collected open-source benchmarks in **PARROT**, highlighting their  
 20 sources, dialect coverage, and key statistics. As shown in Table 2, the collected benchmarks showcase  
 21 a significant diversity with different complexities (e.g., the number of supported dialects).

Table 2: Details of Collected Benchmarks in **PARROT**.

Benchmark	Year	SQL Dialects Supported	Language	Domain Type	Turn	Collection
ATIS	1994	SQLite, MySQL	English	Single-domain	Single	Manual
GeoQuery	1996	MySQL, SQLite	English	Single-domain	Single	Manual
Restaurants	2000	SQLite	English	Single-domain	Single	Manual
Academic	2014	<i>Unspecified</i>	English	Single-domain	Single	Manual
IMDb	2017	<i>Unspecified</i>	English	Single-domain	Single	Manual
Yelp	2017	<i>Unspecified</i>	English	Single-domain	Single	Manual
Scholar	2017	<i>Unspecified</i>	English	Single-domain	Single	Manual
WikiSQL	2017	SQLite3	English	Cross-domain	Single	Manual
Advising	2018	SQLite, MySQL	English	Single-domain	Single	Manual
Spider	2018	SQLite	English	Cross-domain	Single	Manual
SParC	2019	SQLite	English	Cross-domain	Multiple	Manual
CoSQL	2019	SQLite	English	Cross-domain	Multiple	Manual
CSpider	2019	SQLite	Chinese	Cross-domain	Single	Manual
MIMICSQL	2020	SQLite	English	Single-domain	Single	Hybrid <sup>†</sup>
SQUALL	2020	SQLite	English	Cross-domain	Single	Manual
FIBEN	2020	Db2, PostgreSQL	English	Single-domain	Single	Manual
ViText2SQL	2020	General SQL	Vietnamese	Cross-domain	Single	Manual
DuSQL	2020	<i>Unspecified</i>	Chinese	Cross-domain	Single	Hybrid <sup>†</sup>
PortugueseSpider	2021	SQLite	Portuguese	Cross-domain	Single	Hybrid <sup>†</sup>
CHASE	2021	SQLite	Chinese	Cross-domain	Multiple	Manual
Spider-Syn	2021	SQLite	English	Cross-domain	Single	Manual
Spider-DK	2021	SQLite	English	Cross-domain	Single	Manual
Spider-Realistic	2021	SQLite	English	Cross-domain	Single	Manual
KaggleDBQA	2021	SQLite	English	Cross-domain	Single	Manual
SEDE	2021	T-SQL	English	Single-domain	Single	Manual
MT-TEQL	2021	SQLite	English	Cross-domain	Single	Automatic
PAUQ	2022	SQLite	Russian	Cross-domain	Single	Manual
knowSQL	2022	<i>Unspecified</i>	Chinese	Cross-domain	Single	Manual
Dr.Spider	2023	SQLite	English	Cross-domain	Single	Hybrid <sup>†</sup>
BIRD	2023	SQLite	English	Cross-domain	Single	Manual
AmbiQT	2023	SQLite	English	Cross-domain	Single	LLM-aided
ScienceBenchmark	2024	General SQL	English	Single-domain	Single	Hybrid <sup>†</sup>
BookSQL	2024	SQLite	English	Single-domain	Single	Manual
Archer	2024	SQLite	English/ Chinese	Cross-domain	Single	Manual
BULL	2024	SQLite	English/ Chinese	Single-domain	Single	Manual
Spider2	2024	SQLite, DuckDB, PostgreSQL	English	Cross-domain	Single	Manual
TPC-H FROID	2018	T-SQL, PostgreSQL	English	Cross-domain	Single	Hybrid <sup>†</sup>
DSB	2021	T-SQL, PostgreSQL	English	Decision Support	Single	Hybrid <sup>†</sup>
TPC-DS	2005	T-SQL, PostgreSQL	English	Decision Support	Single	Hybrid <sup>†</sup>

Continued on next page

Table 2 – continued from previous page

Benchmark	Year	SQL Dialects Supported	Language	Domain Type	Turn	Collection
SQL-ProcBench	2021	SQL Server, PostgreSQL, IBM Db2	English	Enterprise workloads	Single	Production-derived

<sup>†</sup> **Hybrid** means the dataset was created using both automatic generation and manual annotation.

#### A.4 Experiment Details

We introduce the prompt design adopted in **PARROT**, which facilitates efficient user interaction and enhances LLM-guided SQL understanding. As shown in Table 3, these prompts are well-structured with clear instructions (e.g., translating SQLs among specified dialects) over the objectives (e.g., adhering to the translation criteria).

Table 3: SQL Annotation System and User Prompt in **PARROT**.

System Prompt
<p><b>## CONTEXT ##</b></p> <p>You are a database expert specializing in various SQL dialects, such as <code>**{src_dialect}**</code> and <code>**{tgt_dialect}**</code>, with a focus on accurately translating SQL queries between these dialects.</p> <p><b>## OBJECTIVE ##</b></p> <p>Your task is to translate the input SQL from <code>**{src_dialect}**</code> into <code>**{tgt_dialect}**</code>, ensuring the following criteria are met:</p> <ol style="list-style-type: none"> <li><b>Grammar Compliance</b>: The translated SQL must strictly adhere to the grammar and conventions of <code>{tgt_dialect}</code> (e.g., correct usage of keywords and functions);</li> <li><b>Functional Consistency</b>: The translated SQL should produce the same results and maintain the same functionality as the input SQL (e.g., same columns and data types).</li> <li><b>Clarity and Efficiency</b>: The translation should be clear and efficient, avoiding unnecessary complexity while achieving the same outcome.</li> </ol> <p>During your translation, please consider the following candidate translation points:</p> <ol style="list-style-type: none"> <li><b>Keywords and Syntax</b>: Ensure <code>{tgt_dialect}</code> supports all the keywords from the input SQL, and that the syntax is correct;</li> <li><b>Built-In Functions</b>: Verify that any built-in functions from <code>{src_dialect}</code> are available in <code>{tgt_dialect}</code>, paying attention to the argument types and the return types;</li> <li><b>Data Types</b>: Ensure that <code>{tgt_dialect}</code> supports the data types used in the input SQL. Address any expressions that require explicit type conversions;</li> <li><b>Incompatibilities</b>: Resolve any other potential incompatibility issues during translation.</li> </ol> <p>This task is crucial, and your successful translation will be recognized and rewarded. Please start by carefully reviewing the input SQL and then proceed with the translation.</p>
User Prompt
<p><b>## INPUT ##</b></p> <p>Please translate the input SQL from <code>**{src_dialect}**</code> to <code>**{tgt_dialect}**</code>. The input SQL is:</p> <pre>““sql {sql} ””</pre> <p><b>## OUTPUT FORMAT ##</b></p> <p>Please return your response without any redundant information, strictly adhering to the following format:</p>

Continued on next page

Table 3 – continued from previous page

---

```

““json
{{
"Answer": "The translated SQL",
"Reasoning": "Your detailed reasoning for the translation steps (clear and succinct, no more
than 200 words)",
"Confidence": "The confidence score about your translation (0 - 1)"
}}
““

## OUTPUT ##

```

---

## 28 A.5 Future Work

29 We introduce **PARROT**, the first benchmark specifically designed to evaluate cross-system SQL  
30 translation. By providing a carefully curated, diverse dataset, a suite of testing cases, and a stan-  
31 dardized evaluation protocol, **PARROT** enables a comprehensive and practical assessment of LLM  
32 performance on system-specific SQL translation tasks.

33 While **PARROT** offers a strong foundation, there remain opportunities for improvement in both  
34 the *dataset* and *solution* aspects. (1) On the *dataset* side, we plan to incorporate SQLs with more  
35 complex and diverse translation types to broaden coverage and increase difficulty, thereby presenting  
36 greater challenges for existing LLMs. (2) On the *solution* side, we aim to explore more advanced  
37 LLM-based approaches, such as fine-tuning specialized models with translation capabilities, as well  
38 as hybrid strategies that combine statistical and rule-based methods.

39 These two aspects are closely intertwined and collectively contribute to advancing the development  
40 of more robust, accurate, and generalizable SQL translation methods across diverse database systems.

## 41 B Datasheet for PARROT

42 In this section, we use the framework of **Datasheets for Datasets** to form a datasheet for **PARROT**,  
43 aiming to document the motivation, composition, collection process, recommended uses, and other  
44 information for our benchmark **PARROT**.

### 45 B.1 Motivation

46 Q1. For what purpose was the dataset created? Was there a specific task in mind?

47 The dataset was created to support the task of Cross-System SQL Translation (i.e., SQL-to-SQL  
48 translation), which involves adapting a query written for one database system into its functionally  
49 equivalent form for another. This task addresses a critical gap in existing benchmarks, which are  
50 typically limited to a single system (e.g., SQLite) and fail to capture the syntactic and semantic diver-  
51 sity of real-world SQL dialects. **PARROT** was specifically designed to enable rigorous evaluation  
52 of system-specific SQL translation, providing diverse and realistic query pairs across 22 database  
53 systems to benchmark and advance LLM performance in this underexplored but practically important  
54 area.

55 Q2. Who created this dataset (e.g., which team, research group) and on behalf of which entity  
56 (e.g., company, institution, organization)?

57 The authors of this paper create the **PARROT**. The authors are from the Shanghai Jiao Tong University,  
58 the Tsinghua University and the ByteHouse Team. Please refer to the author list for more details.

59 Q3. Who funded the creation of the dataset?

60 The creation of **PARROT** was supported by the ByteHouse Team. Moreover, it was supported by  
61 National Key R&D Program of China (2023YFB4503600), NSF of China (62232009, 62102215),  
62 Shenzhen Project (CJGJZD20230724093403007), Zhongguancun Lab, Huawei, and Beijing National  
63 Research Center for Information Science and Technology (BNRist).

64 Q4. Any other comments?

65 No.

### 66 B.2 Composition

67 Q5. What do the instances that comprise the dataset represent (e.g., documents, photos, people,  
68 countries)?

69 The instances in the dataset represent pairs of SQL queries, where each pair consists of a source  
70 SQL query written in one database dialect and its functionally equivalent translation in a different  
71 target dialect. These pairs are drawn from real-world applications and open-source benchmarks and  
72 are designed to reflect the syntactic and semantic variations across database systems. Each instance  
73 captures a specific translation challenge arising from differences in SQL syntax, data types, built-in  
74 functions, or system-specific conventions.

75 Q6. How many instances are there in total (of each type, if appropriate)?

76 The dataset includes multiple variants, each containing a different number of SQL-to-SQL translation  
77 pairs: **PARROT (core set)**: 598 carefully curated translation pairs sourced from 38 open-source  
78 benchmarks and real-world business services. **PARROT-Diverse**: 28,003 translation pairs selected  
79 to provide structural and semantic diversity across a wide range of query types. **PARROT-Simple**:  
80 5,306 translation pairs focusing on specialized test cases with simpler query patterns. In total, the  
81 dataset provides **33,907 SQL-to-SQL translation pairs** spanning **22 different database systems**.

82 Q7. Does the dataset contain all possible instances or is it a sample (not necessarily random) of  
83 instances from a larger set?

84 **PARROT** is a newly curated benchmark.

85 Q8. What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or  
86 features?

87 Each instance consists of *raw SQL queries* in textual form, specifically a pair of functionally equivalent  
88 SQL statements: one in the source dialect and one in the target dialect. These queries are unprocessed  
89 and preserve dialect-specific syntax, functions, and data types to reflect real-world translation  
90 scenarios. No engineered features or intermediate representations are included; the dataset is intended  
91 to serve as direct input/output examples for training or evaluating SQL translation systems.

92 **Q9. Is there a label or target associated with each instance?**

93 Yes, the label or target for each instance is the ground truth.

94 **Q10. Is any information missing from individual instances?**

95 No.

96 **Q11. Are relationships between individual instances made explicit (e.g., users' movie ratings,  
97 social network links)?**

98 Yes.

99 **Q12. Are there recommended data splits (e.g., training, development/validation, testing)?**

100 Yes.

101 **Q13. Are there any errors, sources of noise, or redundancies in the dataset?**

102 To ensure data quality, the dataset undergoes a rigorous curation and anonymization process, including  
103 the removal of redundant queries, correction of syntactic inconsistencies, and validation of functional  
104 equivalence across dialects. However, given the complexity and diversity of SQL dialects, minor  
105 inconsistencies may still arise due to edge cases in dialect-specific behavior (e.g., implicit type  
106 casting or null handling). Such cases are carefully reviewed, and filtering mechanisms are applied  
107 to minimize noise and ensure that retained instances reflect meaningful and challenging translation  
108 scenarios.

109 **Q14. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g.,  
110 websites, tweets, other datasets)?**

111 **PARROT** is self-contained.

112 **Q15. Does the dataset contain data that might be considered confidential (e.g., data that is  
113 protected by legal privilege or by doctor-patient confidentiality, data that includes the content  
114 of individuals non-public communications)?**

115 No.

116 **Q16. Does the dataset contain data that, if viewed directly, might be offensive, insulting,  
117 threatening, or might otherwise cause anxiety?**

118 No.

119 **Q17. Does the dataset relate to people?**

120 No.

121 **Q18. Does the dataset identify any subpopulations (e.g., by age, gender)?**

122 No.

123 **Q19. Is it possible to identify one or more natural persons, either directly or indirectly (i.e., in  
124 combination with other data) from the dataset?**

125 No.

126 **Q20. Does the dataset contain data that might be considered sensitive in any way (e.g., data  
127 that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or  
128 union memberships, or locations; financial or health data; biometric or genetic data; forms of  
129 government identification, such as social security numbers; criminal history)?**

130 No.

131 **Q21. Any other comments?**

132 No.

133 **B.3 Collection Process**

134 **Q22. How was the data associated with each instance acquired?**

135 We have elaborated how we construct **PARROT** in the paper.

136 **Q23. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus**  
137 **or sensor, manual human curation, software program, software API)?**

138 We have elaborated how we construct **PARROT** in the paper.

139 **Q24. If the dataset is a sample from a larger set, what was the sampling strategy?**

140 Please refer to the paper for more details.

141 **Q25. Who was involved in data collection process (e.g., students, crowd-workers, contractors)**  
142 **and how were they compensated (e.g., how much were crowdworkers paid)?**

143 The authors of this paper were involved in curating, processing, and reviewing the dataset, and no  
144 compensation was provided.

145 **Q26. Over what timeframe was the data collected? Does this timeframe match the creation**  
146 **timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?**

147 Between June 2024 to May 2025.

148 **Q27. Were any ethical review processes conducted (e.g., by an institutional review board)?**

149 Not applicable. **PARROT** is a synthesized benchmark containing no sensitive information.

150 **Q28. Does the dataset relate to people?**

151 No.

152 **Q29. Did you collect the data from the individuals in question directly, or obtain it via third**  
153 **parties or other sources (e.g., websites)?**

154 Not applicable.

155 **Q30. Were the individuals in question notified about the data collection?**

156 Not applicable.

157 **Q31. Did the individuals in question consent to the collection and use of their data?**

158 Not applicable.

159 **Q32. If consent was obtained, were the consenting individuals provided with a mechanism to**  
160 **revoke their consent in the future or for certain uses?**

161 Not applicable.

162 **Q33. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a**  
163 **data protection impact analysis) been conducted?**

164 Not applicable.

165 **Q34. Any other comments?**

166 No.

167 **B.4 Preprocessing, Cleaning, and/or Labeling**

168 **Q35. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing,**  
169 **tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing**  
170 **of missing values)?**

171 Yes. Please refer to the corresponding section in the paper.

172 **Q36. Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to**  
173 **support unanticipated future uses)?**

174 Yes. Raw data can be found in our GitHub repository.

175 **Q37. Is the software used to preprocess/clean/label the instances available?**  
 176 Yes. We have provided scripts for certain data-cleaning, processing, and extraction tasks in our GitHub  
 177 repository. This repository is open-source and accessible to facilitate potential further research.

178 **Q38. Any other comments?**  
 179 No.

180 **B.5 Uses**

181 **Q39. Has the dataset been used for any tasks already?**  
 182 **PARROT** is newly proposed by us, please refer to the paper for our motivation and usage.

183 **Q40. Is there a repository that links to any or all papers or systems that use the dataset?**  
 184 Currently, no.

185 **Q41. What (other) tasks could the dataset be used for?**  
 186 Please refer to the paper for the motivation of this paper.

187 **Q42. Is there anything about the composition of the dataset or the way it was collected and**  
 188 **preprocessed/cleaned/labeled that might impact future uses?**  
 189 No.

190 **Q43. Are there any tasks for which the dataset should not be used?**  
 191 No.

192 **Q44. Any other comments?**  
 193 No.

194 **B.6 Distribution**

195 **Q45. Will the dataset be distributed to third parties outside of the entity (e.g., company,**  
 196 **institution, organization) on behalf of which the dataset was created?**  
 197 Yes, **PARROT** is open-source under GPL-3 license.

198 **Q46. How will the dataset be distributed (e.g., tarball on website, API, GitHub)**  
 199 We make our source code and data public on the GitHub.

200 **Q47. When will the dataset be distributed?.**  
 201 Already available.

202 **Q48. Will the dataset be distributed under a copyright or other intellectual property (IP) license,**  
 203 **and/or under applicable terms of use (ToU)?**  
 204 Yes. It is under GPL-3 license.

205 **Q49. Have any third parties imposed IP-based or other restrictions on the data associated with**  
 206 **the instances?**  
 207 No.

208 **Q50. Do any export controls or other regulatory restrictions apply to the dataset or to individual**  
 209 **instances?**  
 210 No.

211 **Q51. Any other comments?**  
 212 No.



213 **B.7 Maintenance**

214 **Q52. Who will be supporting/hosting/maintaining the dataset?**

215 The authors is maintaining the dataset.

216 **Q53. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

217 By email or raising issues in GitHub repository.

218 **Q54. Is there an erratum?**

219 Currently, no. If necessary, possible erratum will be released in the README file in GitHub  
220 repository.

221 **Q55. Will the dataset *be updated* (e.g., to correct labeling errors, add new instances, delete  
222 instances)? If so, please describe how often, by whom, and how updates will be communicated  
223 to users (e.g., mailing list, GitHub)?**

224 We will maintain the **PARROT** in the following two years, and we may expand **PARROT** with  
225 increased scale and broader coverage if necessary. Relevant information will be released in the  
226 GitHub documentation if there are any updates.

227 **Q56. If the dataset relates to people, are there applicable limits on the retention of the data  
228 associated with the instances (e.g., were individuals in question told that their data would be  
229 retained for a fixed period of time and then deleted)?**

230 Not applicable.

231 **Q57. Will older versions of the dataset continue to be supported/hosted/maintained?**

232 Not applicable.

233 **Q58. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism  
234 for them to do so?**

235 Yes, as long as follow the license, further contributions will be warmly welcomed.

236 **Q59. Any other comments?**

237 No.