Additional information

ACM Reference Format:
. 2024. Additional information. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Criteria for categories of datasets analysis aspects

The categorization of the results for each aspect is made based on the following criteria:

1.1 SQL queries

- Structural variety
  - Low: At least 50% of the dataset contains queries of type "select-from" or "select-from-where".
  - Medium: Structural combinations that contain at least one of the order by, group by having, or limit clauses exist in a percentage higher than 50%.
  - High: At least 40% of the queries contain nestings and the average depth in these queries is more than 2.
- Operators complexity
  - Low: Combinations of only joins, aggregates, logical and comparison operators comprise more than 75% of the dataset.
  - Medium: Combinations containing any of the rest of the operator types constitute more than 25% of the dataset.
  - High: Combinations containing any of the rest of the operator types constitute more than 50% of the dataset.
- Operators usage:
  - Low: The average number of operators is smaller than 5.
  - Medium: The average number of operators is between 5 and 15.
  - High: The average number of operators is more than 15.
- Schema usage
  - Low: Less than 5 columns or 3 tables are used in at least 75% of the queries in the dataset.
  - Medium: 5-10 columns or 3-5 tables are used in more than 25% of the dataset.
  - High: More than 10 columns or more than 5 tables are used in more than 25% dataset.
- Content usage:
  - At most 2 values are used in at least 50% of the dataset.
  - 3-10 values are used in at least 50% of the dataset.
  - More than 10 values are used in at least 25% of the dataset.

1.2 Natural language questions

- Complexity
  - Low: The average length of the questions is lower than 50 characters.
  - Medium: The average length of the questions are between 50 and 80
  - High: The average length of the questions is higher than 80
- Schema usage
  - Low: The average percentage of exact schema references is below 20%
  - Medium: The average percentage of exact schema references is between 20% - 50%
  - High: The average percentage of exact schema references is higher than 50%

1.3 Databases

- Schema complexity
  - Low: The average number of tables in the databases is less or equal to 5 or the average number of columns in the databases is less than 100.
  - Medium: The average number of tables in the databases is between 6 and 25 or the average number of columns in the databases is between 100 and 1000.
  - High: The average number of tables in the databases is more than 25 or the average number of columns in the databases is more than 1000.
- Schema quality
  - Low: The are databases with percentage of explainable schema elements less than 50.
  - Medium: The percentage of explainable schema elements is more than 50 in all databases.
  - High: The percentage of explainable schema elements is more than 80 in all databases.
- Database size
  - Low: There are databases that contain less than 1000 average table rows.
  - Medium: All the databases contain more than 1000 average table rows and less than 1 million average table rows.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

– High: There are databases containing more than 1 million average table rows.

2 Datasets Analysis figures

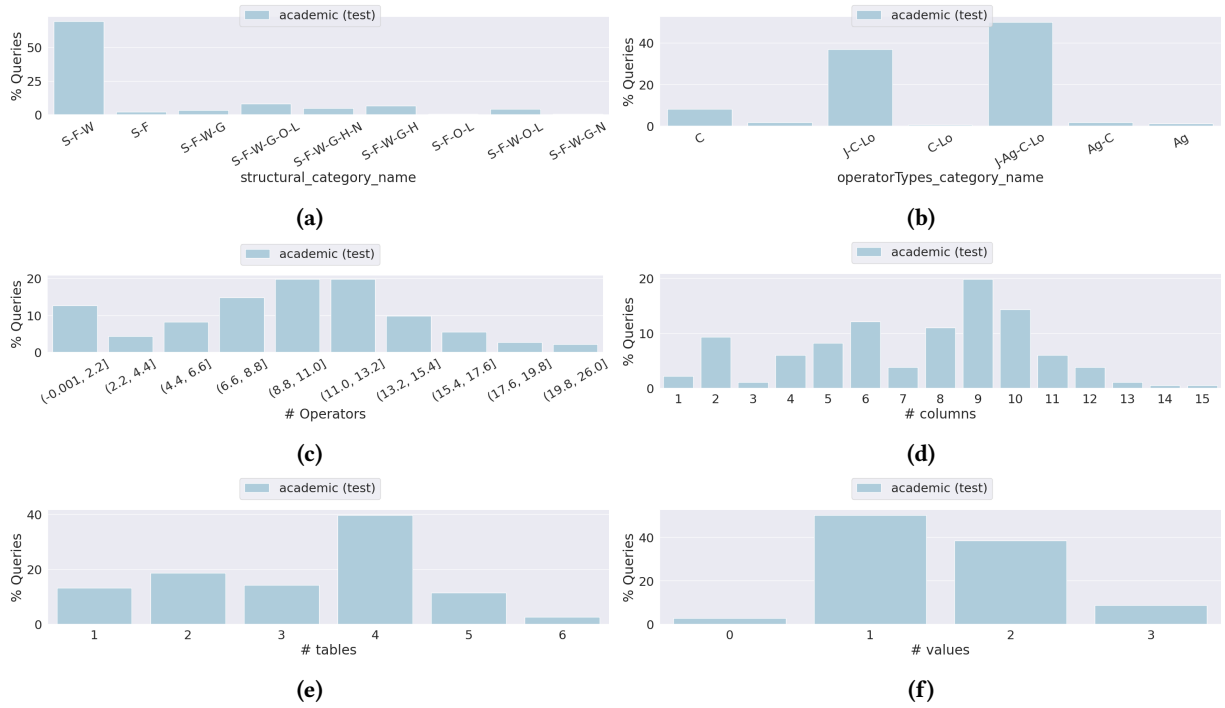


Figure 1. Statistics of Academic sql queries

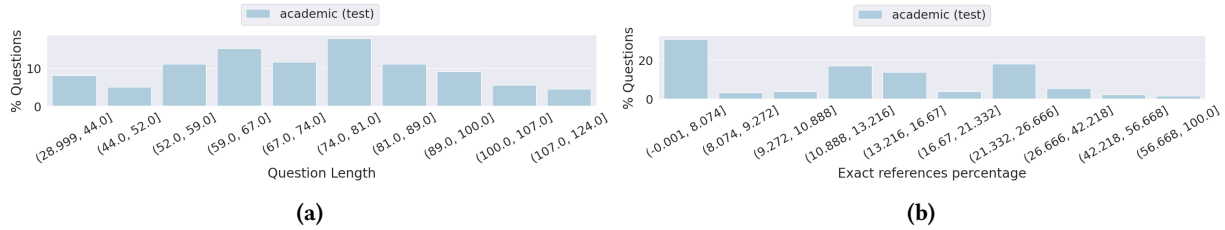


Figure 2. Statistics of Academic natural language questions

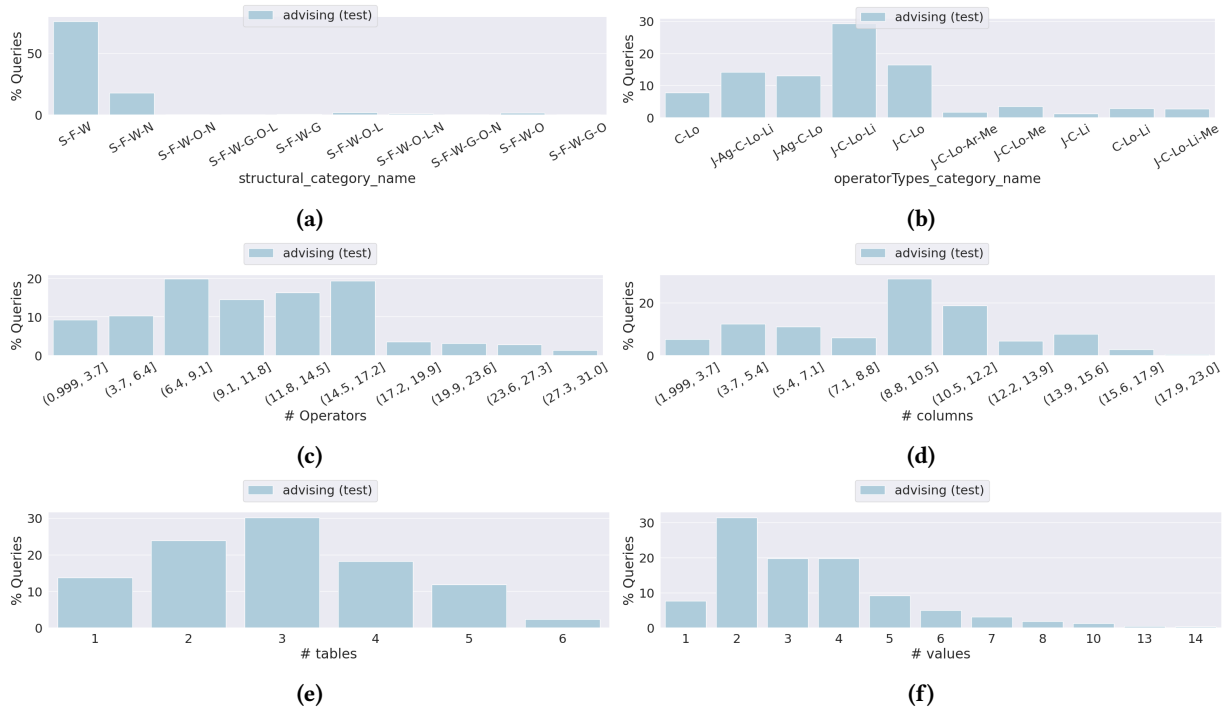


Figure 3. Statistics of Advising sql queries

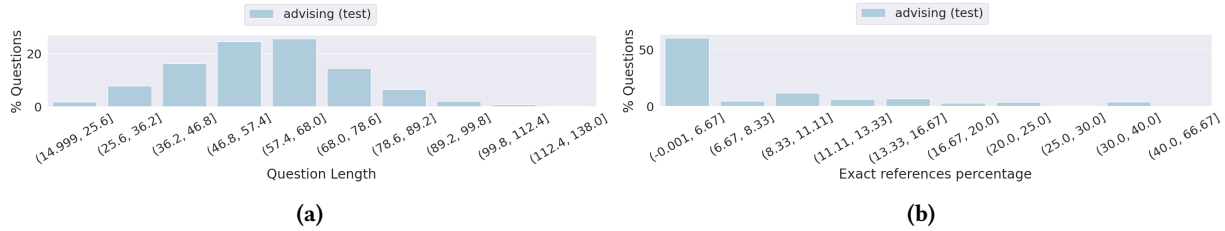
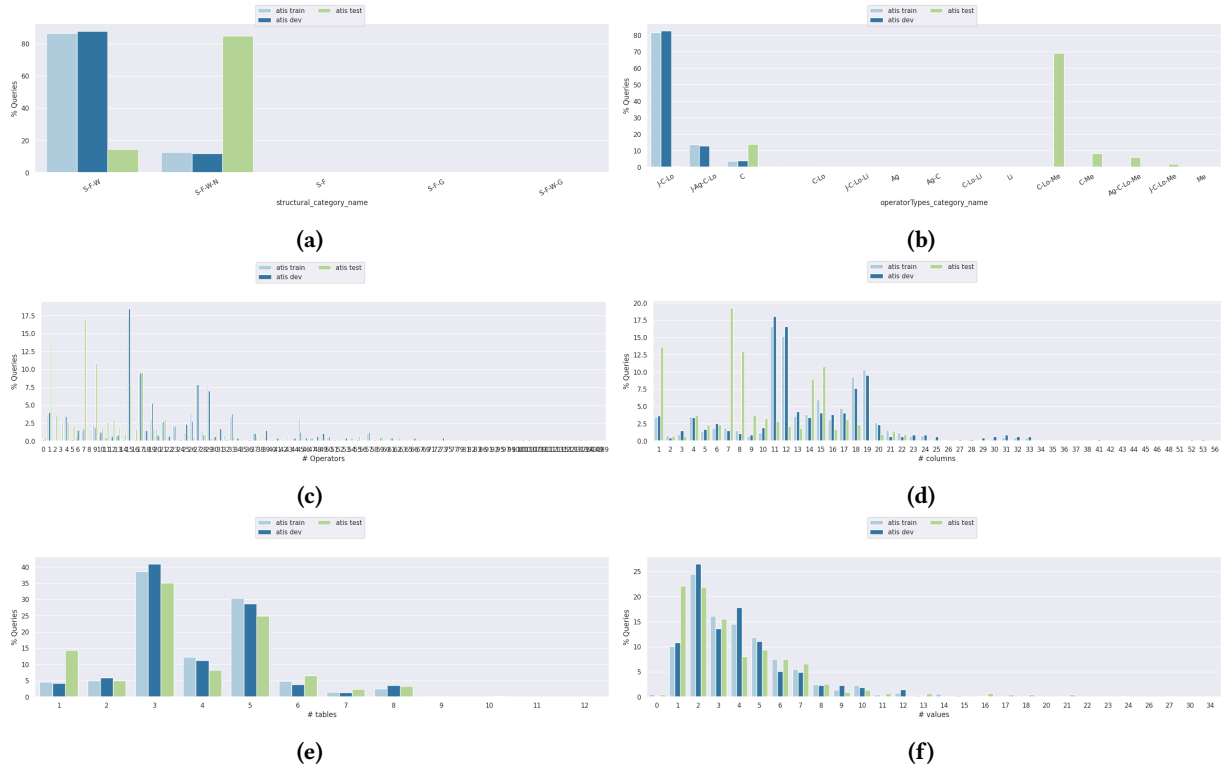
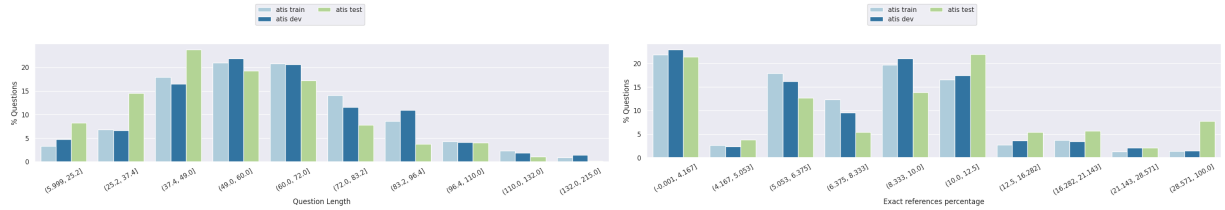


Figure 4. Statistics of Advising natural language questions



(a) (b) (c) (d) (e) (f)

Figure 5. Statistics of Atis sql queries



(a) (b)

Figure 6. Statistics of Atis natural language questions

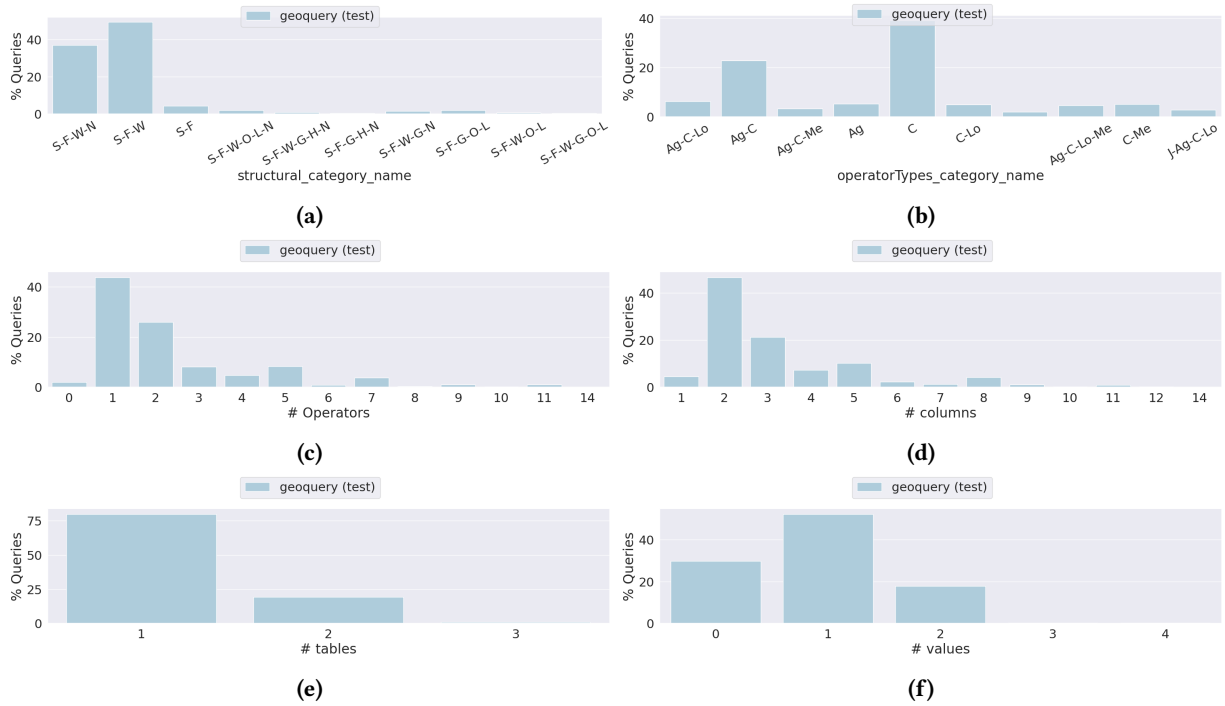


Figure 7. Statistics of Geoquery sql queries

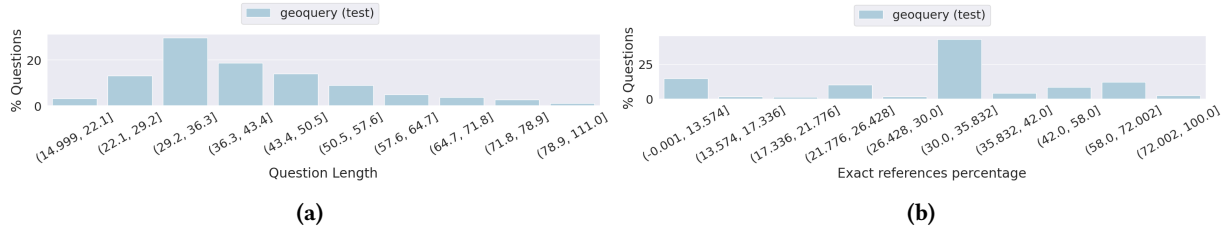


Figure 8. Statistics of Geoquery natural language questions

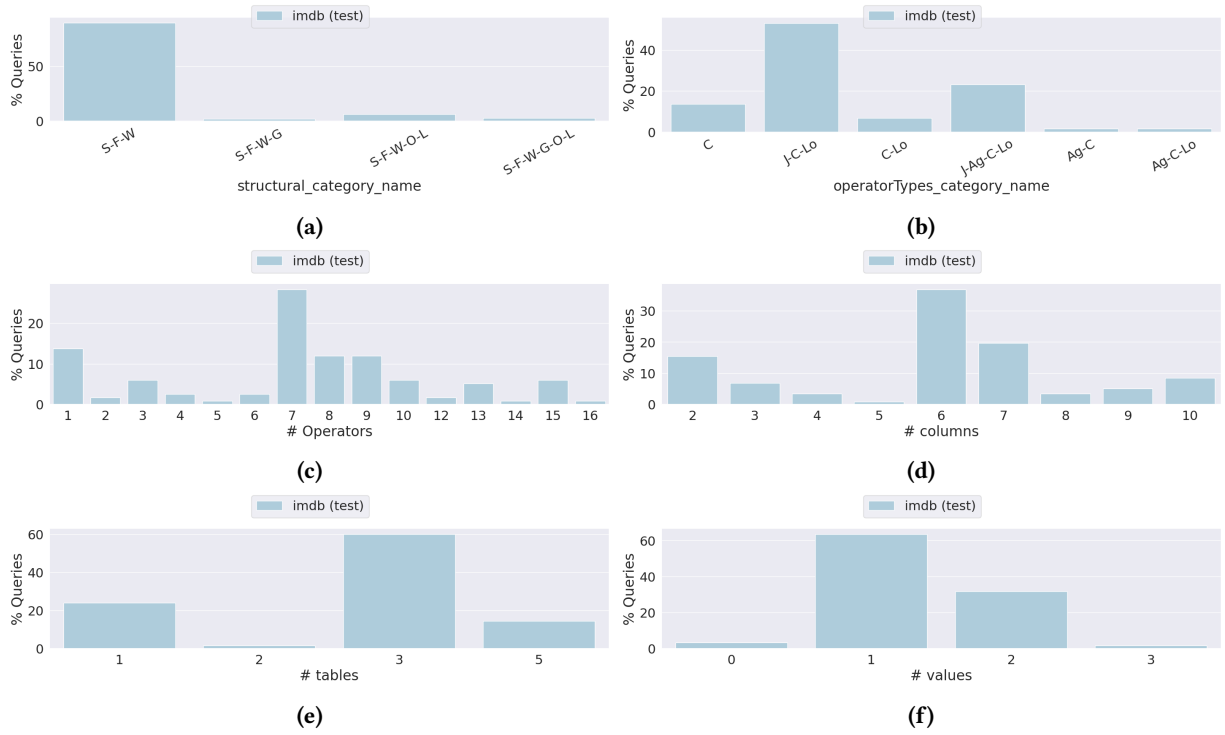


Figure 9. Statistics of Imdb sql queries

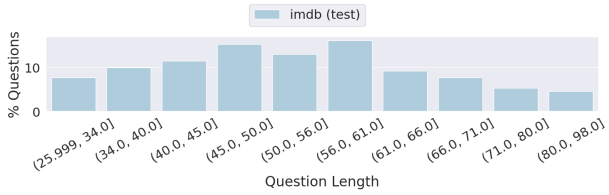


Figure 10. Statistics of Imdb natural language questions

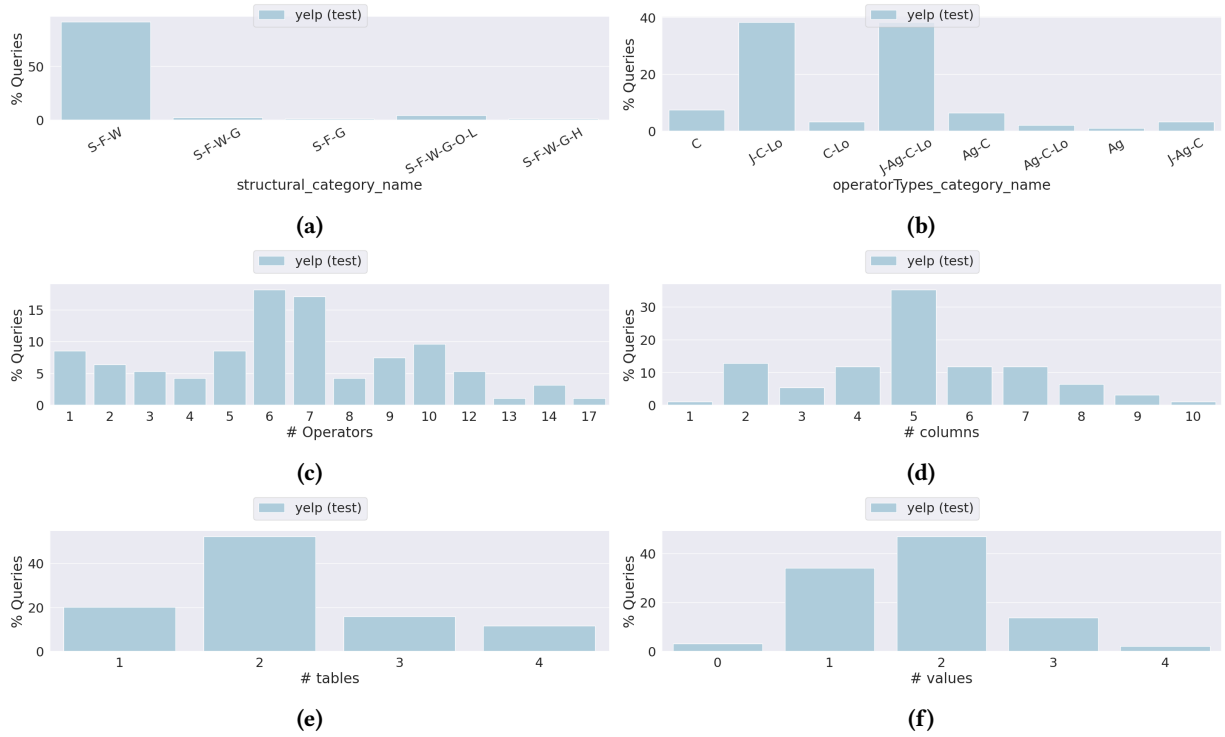


Figure 11. Statistics of Yelp sql queries

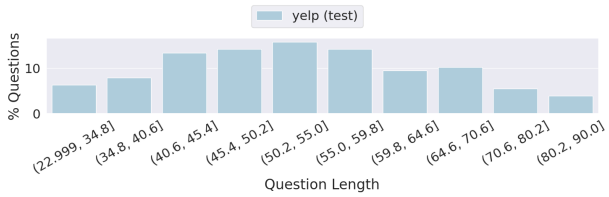


Figure 12. Statistics of Yelp natural language questions

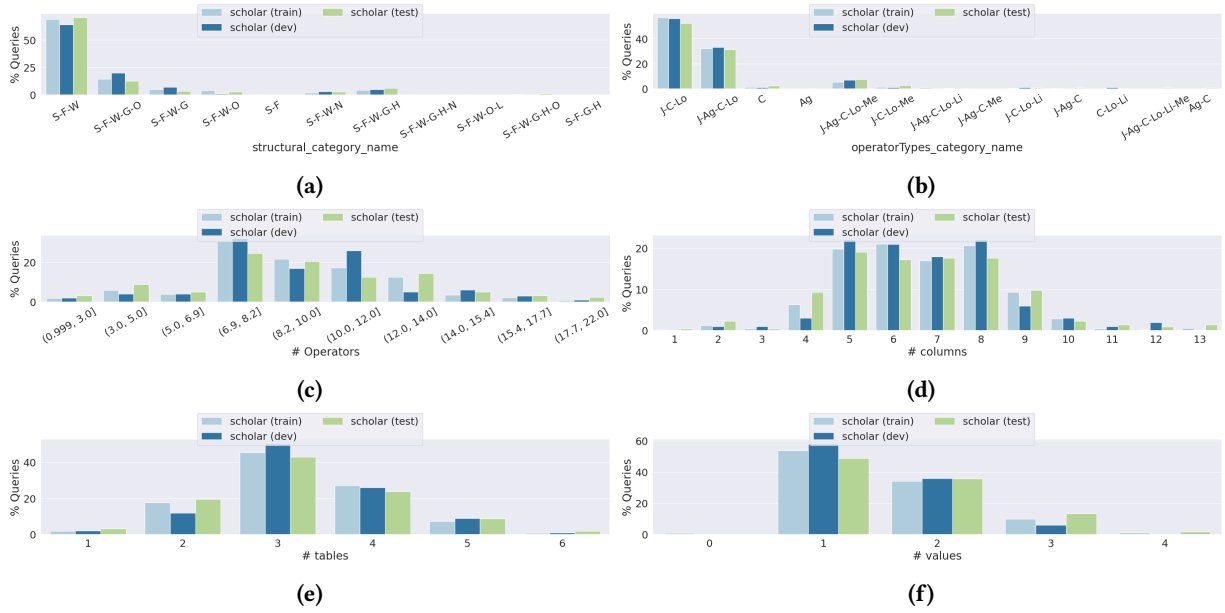


Figure 13. Statistics of Scholar sql queries

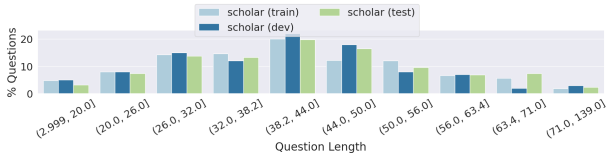


Figure 14. Statistics of Scholar natural language questions

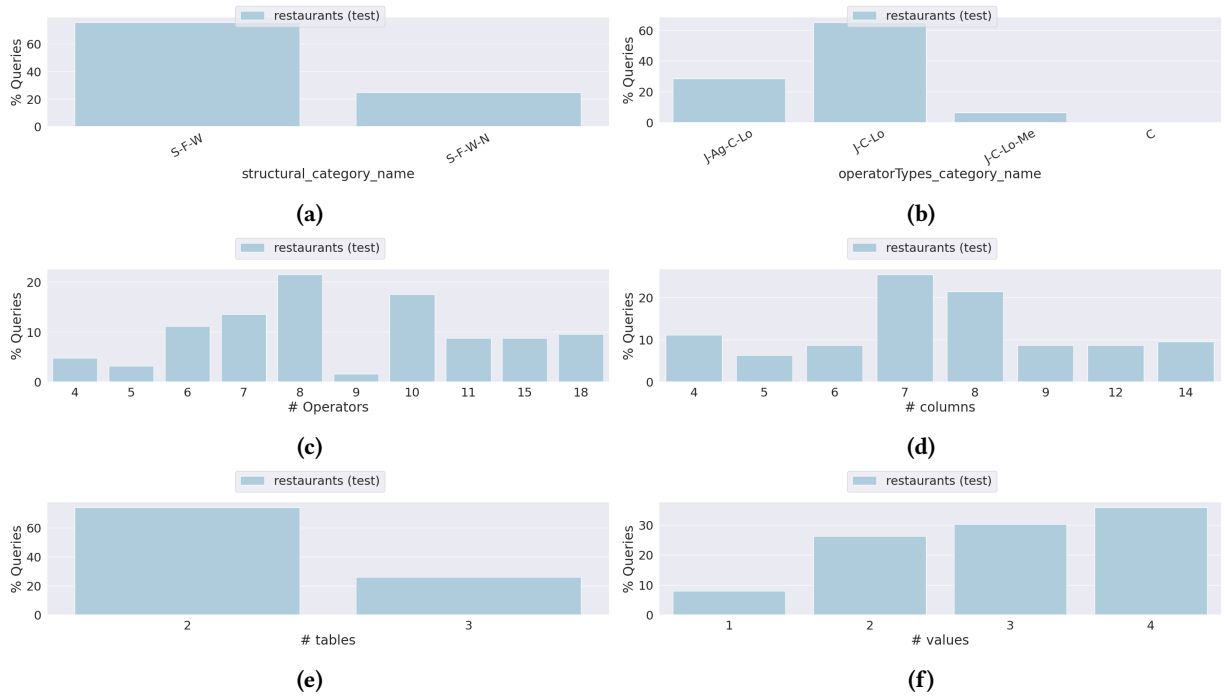


Figure 15. Statistics of Restaurants sql queries

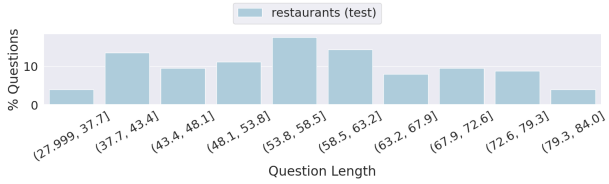


Figure 16. Statistics of Restaurants natural language questions

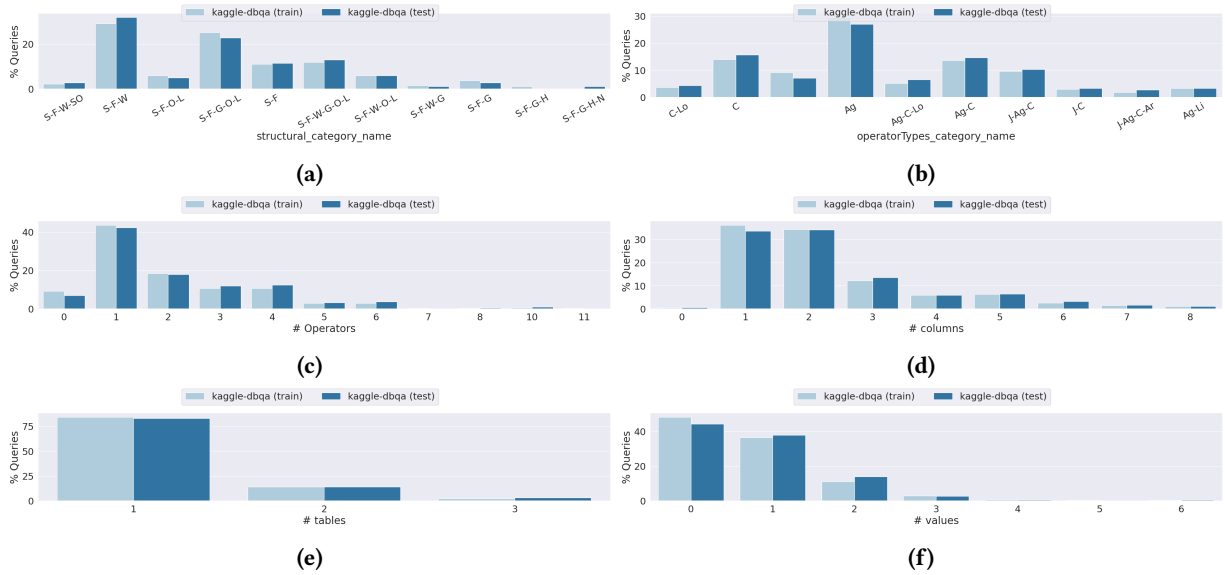


Figure 17. Statistics of KaggleDBQA sql queries

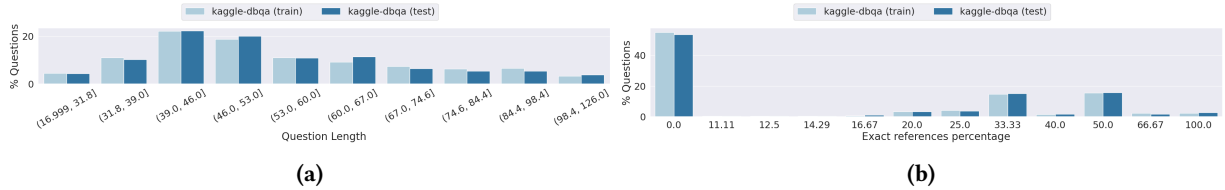


Figure 18. Statistics of KaggleDBQA natural language questions