Efficient Data Wrangling with LLMs using Code Generation

Xue Li [1], Till Döhmen [2]

[1] University of Amsterdam, Amsterdam, The Netherlands
x.li3@uva.nl
[2] MotherDuck, Amsterdam, The Netherlands
till@motherduck.com

Abstract While LLM-based data wrangling approaches that process each row of data have shown promising benchmark results[1], computational costs still limit their suitability for real-world use cases on large datasets. We revisit code generation[2, 3] using LLMs for various data wrangling tasks, which show promising results particularly for data transformation tasks (up to 37.2 points improvement on F1 score) at much lower computational costs. We furthermore identify shortcomings of code generation methods especially for semantically challenging tasks, and consequently envision an approach[4] that combines program generation with a routing mechanism using LLMs. The envisioned approach involves a shift from employing LLMs on a per-row basis to prompting LLMs for function generation that addresses the bulk of records. The approach includes a task router that divides tasks from code-solvable to not code-solvable, and a data router that distinguishes data from code-applicable to not code-applicable. Further research will experiment with each component of the proposed workflow, exploring *when* it is appropriate to use LLMs for data wrangling.

References

[1] Avanika Narayan, Ines Chami, Laurel Orr, and Christopher Ré. 2022. Can Foun- dation Models Wrangle Your Data? Proceedings of the VLDB Endowment 16, 4 (2022), 738–746.

[2] José Cambronero, Sumit Gulwani, Vu Le, Daniel Perelman, Arjun Radhakrishna, Clint Simon, and Ashish Tiwari. 2023. Flashfill++: Scaling programming by example by cutting to the chase. Proceedings of the ACM on Programming Languages 7, POPL (2023), 952–981.

[3] Yeye He, Xu Chu, Kris Ganjam, Yudian Zheng, Vivek Narasayya, and Surajit Chaudhuri. 2018. Transform-data-by-example (TDE): an extensible search engine for data transformations. Proc. VLDB Endow. 11, 10 (jun 2018), 1165–1177. https://doi.org/10.14778/3231751.3231766

[4] Xue Li and Till Döhmen. Towards efficient data wrangling with llms using code generation. In Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning, DEEM '24, pp. 62–66.