

Seminararbeit AI-assisted programming and data analysis

Data Profiling with LLM's - A Case Study on HR Data

Vincent-Konstantin Kapp

Prof. Dr. Martin Spindler

31. October 2024

Abstract

Datengetriebene Prozesse können Unternehmen nicht nur dabei helfen, ihre Prozesse zu optimieren, sondern auch bei der Entwicklung neuer Strategien. Um festzustellen, welche Daten in den Datenbanken vorliegen und wie diese zueinander stehen, braucht es Mitarbeiter*innen, die einzeln überprüfen, welche Daten vorhanden sind und in welchen Zusammenhängen sie basierend auf Common Sense existieren. Data Profiling stellt Unternehmen mit historisch gewachsenen Datenbanksystemen vor Herausforderungen für die datengestützte Transformation. Können die aktuellen Entwicklungen von Large Language Modellen (LLMs) dazu beitragen, automatisiert die Zusammenhänge von Spalten zu erkennen? Um diese Frage zu beantworten, widmet sich diese Seminararbeit, aufbauend auf der Arbeit von Trummer (2024), der Untersuchung, wie gut LLMs anhand einer Case Study die Korrelation von Daten erkennen können.

1 Einleitung

2 Hintergrund

2.1 Data Profiling

2.2 Korrelationserkennung in Datenbanken

(Trummer 2023).

2.3 Sprachmodelle und deren Anwendung in der Datenanalyse

(Arora et al. 2023).

3 Methodik der Fallstudie

3.1 Datenbasis und Vorbereitung

3.2 Vorgehensweise zur Korrelationsanalyse mit LLM

3.3 Benchmark und Bewertungsmetriken

4 Durchführung der Fallstudie: Gender-Pay-Gap-Analyse

Meine Überlegung ist es das Ergebnisse von (Bach, Chernozhukov, and Spindler 2024).

4.1 Fragestellung

4.2 Implementierung des LLM für Korrelationserkennung

4.3 Benchmarking-Ergebnisse

5 Vergleich und Auswertung der Ergebnisse

5.1 Vergleich der Ergebnisse mit Prompoting

5.2 Szenariobasierte Auswertung

5.3 Bewertung der Korrelationserkennung

6 Diskussion der Ergebnisse und Implikationen

6.1 Interpretation der Ergebnisse

6.2 Limitationen

6.3 Praktische Anwendungsmöglichkeit

7 Fazit und Ausblick

References

- Arora, Simran, Brandon Yang, Sabri Eyuboglu, Avanika Narayan, Andrew Hojel, Immanuel Trummer, and Christopher Ré. 2023. “Language Models Enable Simple Systems for Generating Structured Views of Heterogeneous Data Lakes.” Journal Article. *arXiv Preprint arXiv:2304.09433*.
- Bach, Philipp, Victor Chernozhukov, and Martin Spindler. 2024. “Heterogeneity in the US Gender Wage Gap.” *Journal of the Royal Statistical Society Series A: Statistics in Society* 187 (1): 209–30.
- Trummer, Immanuel. 2023. “Can Large Language Models Predict Data Correlations from Column Names?” Journal Article. *Proceedings of the VLDB Endowment* 16 (13): 4310–23.