Metadata Matters in Dense Table Retrieval

Daniel Gomm and Madelon Hulsebos

Centrum Wiskunde & Informatica (CWI), Amsterdam, Netherlands
{daniel.gomm,madelon}@cwi.nl

1 Introduction

Recent advances in Large Language Models (LLMs) have enabled powerful systems that perform tasks by reasoning over tabular data [9, 10, 13, 7, 4]. While these systems typically assume relevant data is provided with a query, real-world use cases are mostly open-domain, meaning they receive a query without context regarding the underlying tables. Retrieving relevant tables is typically done over dense embeddings of serialized tables [5]. Yet, there is limited understanding of the effectiveness of different inputs and serialization methods for using such off-the-shelf text-embedding models for table retrieval. In this work, we show that different serialization strategies result in significant variations in retrieval performance. Additionally, we surface shortcomings in commonly used benchmarks applied in open-domain settings, motivating further study and refinement.

2 How to construct table embeddings?

There exist many degrees of freedom in deciding on the parameters for transforming tabular data into a 1-dimensional sequence to extract the embedding. This sequence may include column names, table content (fully or partially), and contextual metadata, formatted as, for example, JSON, Markdown, or HTML. Additionally, row sampling strategies must be defined when including tables only partially. We use the TARGET benchmark [5] to systematically evaluate the influence of the different parameters across various datasets for tabular reasoning. We study the impact of limiting the number of rows included in the sequence, randomly sampling rows from tables, and serializing them row-wise in JSON format. Our experiments reveal significant variations in retrieval performance across parameter settings and models.

In the FeTaQA dataset [7], tables are associated with page and section titles as metadata. Figure 1 shows that this contextual metadata is highly relevant for retrieval. Including this metadata in the embedded sequence improves the average recall@3 by 0.42, showing the importance of contextualizing tables for retrieval. The interplay between metadata inclusion and the number of sampled rows varies by model capability. While stronger models like *gte-large-en-v1.5* benefit from adding rows alongside metadata (notable on FeTaQA and OTT-QA [2]), others show performance degradation, suggesting the loss of contextual metadata information in the embeddings. This is best exemplified by the results

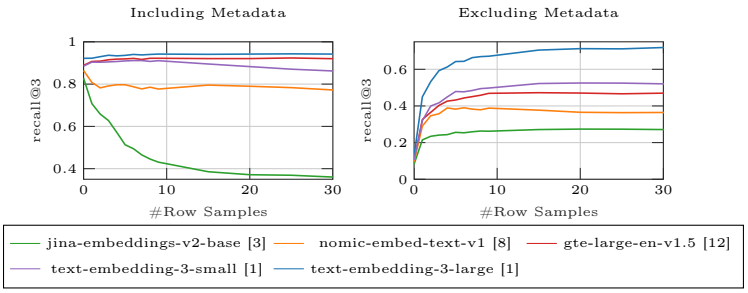

Fig. 1. Recall@3 for different numbers of row samples on the FeTaQA dataset. The left plot shows results when including metadata, and the right one when excluding it.

of *jina-embeddings-v2-base* where the retrieval performance drastically decays the more rows are included. In settings where no metadata is included, embedding more rows improves retrieval with diminishing returns. Most models show little improvement when including more than 10 rows in this setting.

Overall, we find that recent long-context text embedding models perform best, achieving their best results when combining a small sample of rows with meaningful metadata. There is not a single best parameter combination that generalizes across all embedding models. Instead, achieving the best performance is only possible by thorough experimentation with each embedding model.

3 Semantic query coverage for open-domain retrieval

During our experiments, we found large differences in how well queries from different datasets transfer to open-domain settings. We refer to queries that themselves contain sufficient information to identify relevant tables, independent of external context, as *self-contained* (e.g., "When did season 6 of Reborn! air?" - FeTaQA). Queries lacking this information are *under-specified* (e.g., "How many exams are there?" - Spider). We classify queries from popular benchmarks using an LLM (gpt-4o-mini). This analysis reveals shortcomings in existing datasets: FeTaQA has 11% under-specified queries, OTT-QA 17%. In comparison, the popular Text-to-SQL datasets BIRD [6] and Spider [11] exhibit even larger proportions of under-specified queries (36% and 54%, respectively).

Conclusion In this study, we show that table serialization strategies and contextual metadata critically impact retrieval performance, with optimal configurations depending on embedding model capabilities. We further highlight limitations in existing benchmarks complicating realistic evaluation. Future work should perform mechanistic studies of how embeddings encode tabular structure and metadata, clarifying how to best capture table semantics. Additionally, the importance of contextual metadata warrants exploring generative methods to synthesize metadata for tables lacking context. Finally, our study of semantic query coverage motivates efforts to create new or adapt existing datasets for robust open-domain evaluation.

References

1. New embedding models and API updates, https://openai.com/index/new-embedding-models-and-api-updates/
2. Chen, W., Chang, M.W., Schlinger, E., Wang, W., Cohen, W.W.: Open Question Answering over Tables and Text (Oct 2020), https://arxiv.org/abs/2010.10439v2
3. Günther, M., Ong, J., Mohr, I., Abdessalem, A., Abel, T., Akram, M.K., Guzman, S., Mastrapas, G., Sturua, S., Wang, B., Werk, M., Wang, N., Xiao, H.: Jina Embeddings 2: 8192-Token General-Purpose Text Embeddings for Long Documents (Feb 2024). https://doi.org/10.48550/arXiv.2310.19923, http://arxiv.org/abs/2310.19923, arXiv:2310.19923 [cs]
4. Herzig, J., Nowak, P.K., Müller, T., Piccinno, F., Eisenschlos, J.: TaPas: Weakly Supervised Table Parsing via Pre-training. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4320–4333. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.398, https://aclanthology.org/2020.acl-main.398/
5. Ji, X., Parameswaran, A., Hulsebos, M.: TARGET: Benchmarking Table Retrieval for Generative Tasks (Oct 2024), https://openreview.net/forum?id=gGGvnjFUfL
6. Li, J., Hui, B., Qu, G., Yang, J., Li, B., Li, B., Wang, B., Qin, B., Geng, R., Huo, N., Zhou, X., Ma, C., Li, G., Chang, K.C., Huang, F., Cheng, R., Li, Y.: Can LLM already serve as a database interface? a big bench for large-scale database grounded text-to-SQLs. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. pp. 42330–42357. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (Dec 2023)
7. Nan, L., Hsieh, C., Mao, Z., Lin, X.V., Verma, N., Zhang, R., Kryściński, W., Schoelkopf, H., Kong, R., Tang, X., Mutuma, M., Rosand, B., Trindade, I., Bandaru, R., Cunningham, J., Xiong, C., Radev, D., Radev, D.: FeTaQA: Free-form Table Question Answering. Transactions of the Association for Computational Linguistics **10**, 35–49 (2022). https://doi.org/10.1162/tacl_a_00446, https://aclanthology.org/2022.tacl-1.3, place: Cambridge, MA Publisher: MIT Press
8. Nussbaum, Z., Morris, J.X., Duderstadt, B., Mulyar, A.: Nomic Embed: Training a Reproducible Long Context Text Embedder (Feb 2024). https://doi.org/10.48550/arXiv.2402.01613, http://arxiv.org/abs/2402.01613, arXiv:2402.01613 [cs]
9. Pourreza, M., Li, H., Sun, R., Chung, Y., Talaei, S., Kakkar, G.T., Gan, Y., Saberi, A., Ozcan, F., Arik, S.O.: CHASE-SQL: Multi-Path Reasoning and Preference Optimized Candidate Selection in Text-to-SQL. In: Proceedings of the 13th International Conference on Learning Representations (2025). https://doi.org/10.48550/arXiv.2410.01943, https://openreview.net/forum?id=CvGqMD5OtX, arXiv:2410.01943 [cs]
10. Qin, B., Hui, B., Wang, L., Yang, M., Li, J., Li, B., Geng, R., Cao, R., Sun, J., Si, L., Huang, F., Li, Y.: A Survey on Text-to-SQL Parsing: Concepts, Methods, and Future Directions (Aug 2022). https://doi.org/10.48550/arXiv.2208.13629, http://arxiv.org/abs/2208.13629, arXiv:2208.13629 [cs]
11. Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., Ma, J., Li, I., Yao, Q., Roman, S., Zhang, Z., Radev, D.: Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task (Feb 2019). https://doi.org/10.48550/arXiv.1809.08887, http://arxiv.org/abs/1809.08887, arXiv:1809.08887 [cs]

12. Zhang, X., Zhang, Y., Long, D., Xie, W., Dai, Z., Tang, J., Lin, H., Yang, B., Xie, P., Huang, F., Zhang, M., Li, W., Zhang, M.: mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval. In: Dernoncourt, F., Preoţiuc-Pietro, D., Shimorina, A. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track. pp. 1393–1412. Association for Computational Linguistics, Miami, Florida, US (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp-industry.103, https://aclanthology.org/2024.emnlp-industry.103/
13. Zhao, Y., Chen, L., Cohan, A., Zhao, C.: TaPERA: Enhancing Faithfulness and Interpretability in Long-Form Table QA by Content Planning and Execution-based Reasoning. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12824–12840. Association for Computational Linguistics, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.acl-long.692, https://aclanthology.org/2024.acl-long.692