# WikiDBs - Supplementary Material

Liane Vogel[1], Jan-Micha Bodensohn[2,1], Carsten Binnig[1,2]
[1]Technical University of Darmstadt, Germany
[2]German Research Center for Artificial Intelligence (DFKI), Darmstadt, Germany

June 12, 2024

## 1 Preliminary Submission of the Dataset

For review purposes, we provide 20,000 relational databases created from Wikidata for download, the full 100,000 will be released along with the published paper.
We chose this approach for several reasons: First, releasing a subset of the corpus to reviewers allows us to incorporate reviewer feedback into the final version of our dataset. The process of creating the full dataset is costly, we spent a total of 375$ in OpenAI API credits to paraphrase the around 17,000 databases in our dataset which have up to twenty tables. We plan to invest the necessary amount of approximately $2,500$$ to paraphrase the entire dataset after incorporating reviewer feedback.
The dataset is currently only available for the reviewers as a download, the final corpus will be published on Zenodo (similar to our WikiDBs10k version, please see Section 1 of our paper) and on Huggingface Datasets. We do not plan to openly release the preliminary version of the 20,000 databases to avoid confusion about multiple versions of the dataset once the WikiDBs dataset with 100,000 databases is published.

## 2 Additional Material

Our supplementary material includes:

- The following download link for the 20,000 relational databases:
`https://drive.google.com/drive/folders/1wMRFro0ydQghmYeavBaBv_IUsPobT_JK?usp=sharing`
- The dataset description of WikiDBs in form of a *Datasheet for datasets* (datasheet.pdf)
- A croissant format json file (croissant.json)
- The following link to our repository: (the code is published under CC BY 4.0 license)
`https://github.com/DataManagementLab/wikidbs-public`

## 3 Responsibility

The authors bear all responsibility in case of violation of rights, etc. We confirm that the dataset is licensed under a Creative Commons Attribution 4.0 International License.