A Supplementary Material

A.1 Dataset Access

The Genomics LRB is hosted publicly as a dataset repository on HuggingFace at https://hf.co/datasets/InstaDeepAI/genomics-long-range-benchmark where both the data and files to process the data can be found. The meta data associated with our benchmark is located in the README on our HuggingFace repository. The croissant meta data tool was not able to be used due to the custom nature of our dataset files.

A.2 Long Term Preservation

The Genomics LRB is hosted on HuggingFace where users may view and download data by accessing the repository through their browser or can use Hugging Face's API to access and download the dataset programmatically, with a single line of code. In addition to ease-of-use, Hugging Face repositories provide a reliable, stable way to store data and is a framework that is consistently maintained.

Users may raise issues and submit pull requests via the Genomic LRB's Hugging Face webpage which will be addressed by the authors.

A.3 Intended Use

The Genomics LRB is intended to be used by the research community as a tool to assess the capabilities of genomic language models to ultimately lead to their improvement.

A.4 License

The Genomics LRB is licensed under the Creative Commons Attribution Non Commercial Share Alike 4.0 license.

A.5 Author Statement

Authors of this paper bear all responsibility in case of violation of rights.