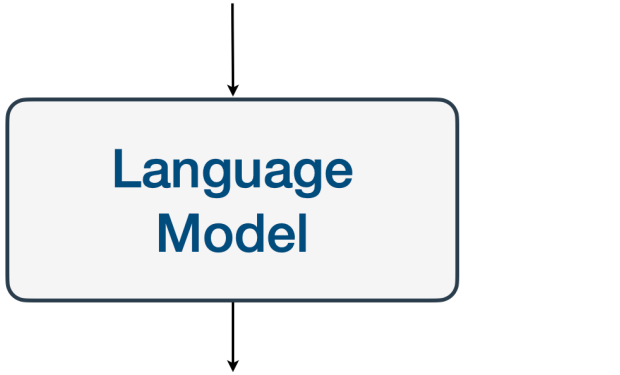
Understanding Transformers for Time Series: Rank Structure, Flow-of-ranks, and Compressibility

Annan Yu, Danielle C. Maddix, Boran Han, Xiyuan Zhang,
Abdul Fatir Ansari, Oleksandr Shchur, Christos Faloutsos,
Andrew Gordon Wilson, Michael W. Mahoney, Yuyang Wang

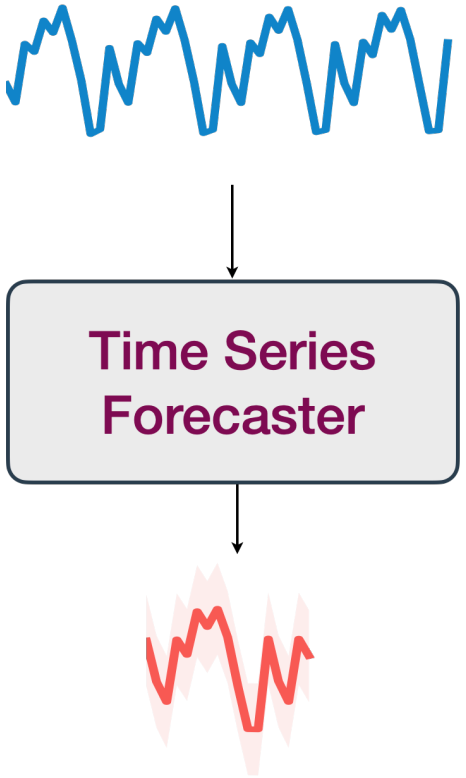
ICLR 2026

Introduction

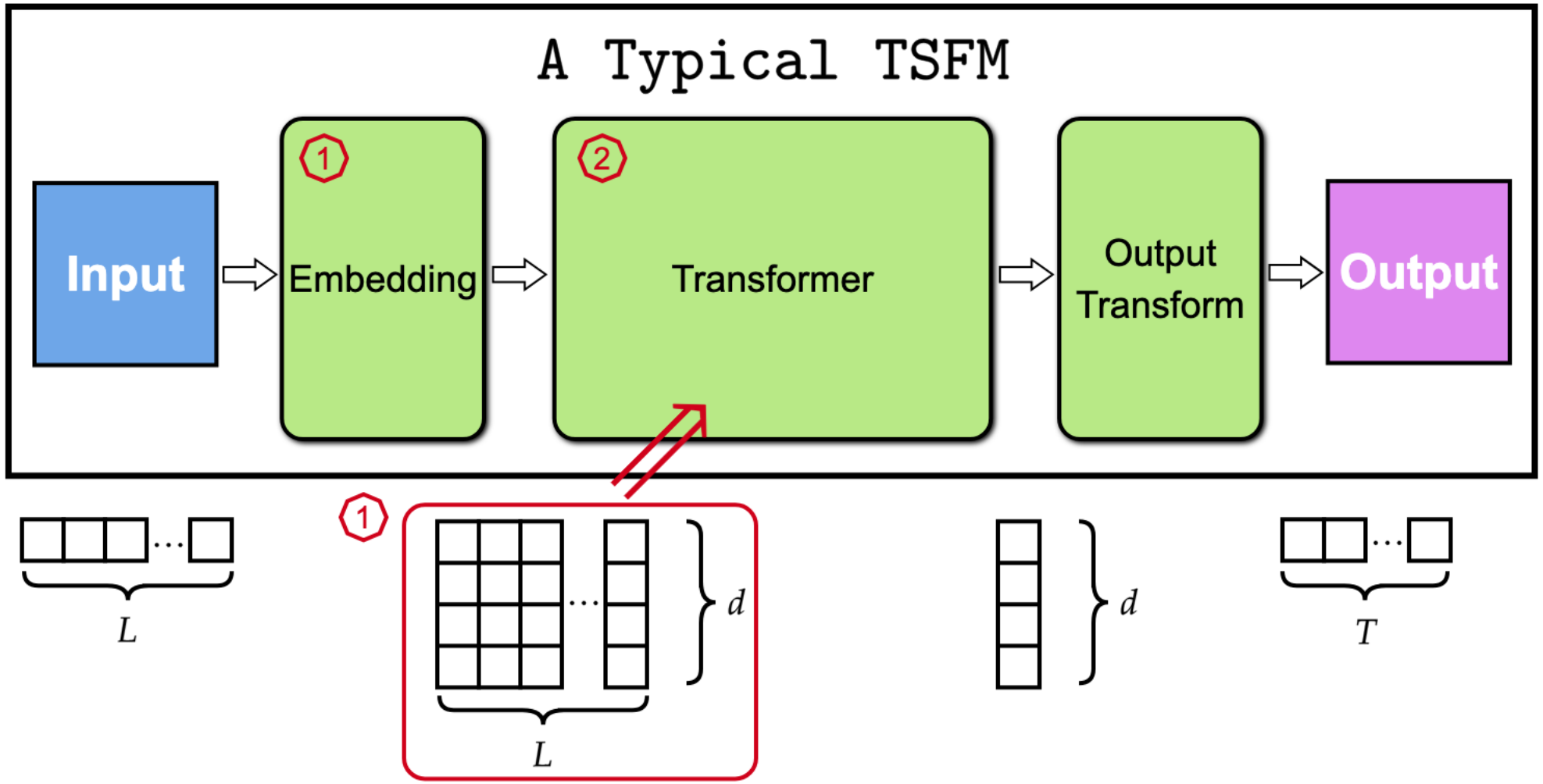
“Three Rings for the Elven-kings under the sky,
Seven for the Dwarf-lords in their halls of stone,
Nine for Mortal Men doomed to die,
One for the Dark Lord on his dark throne
In the Land of Mordor where the Shadows lie.”



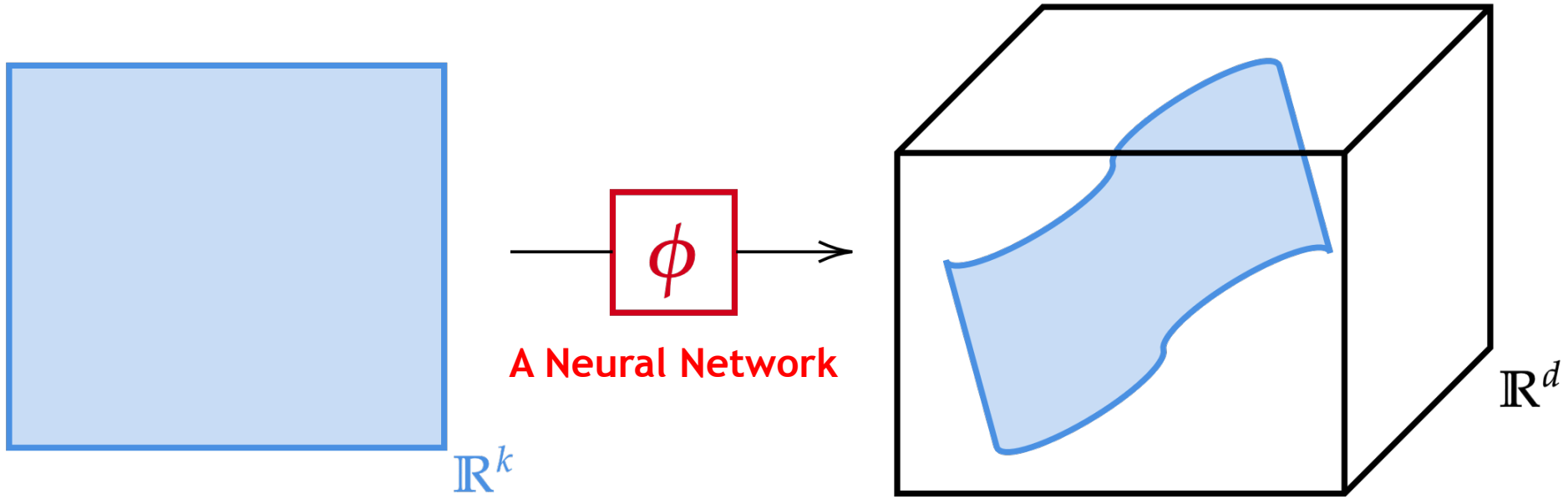
“One Ring to rule them all, One Ring to find them,
One Ring to bring them all and in the darkness bind them
In the Land of Mordor where the Shadows lie.”



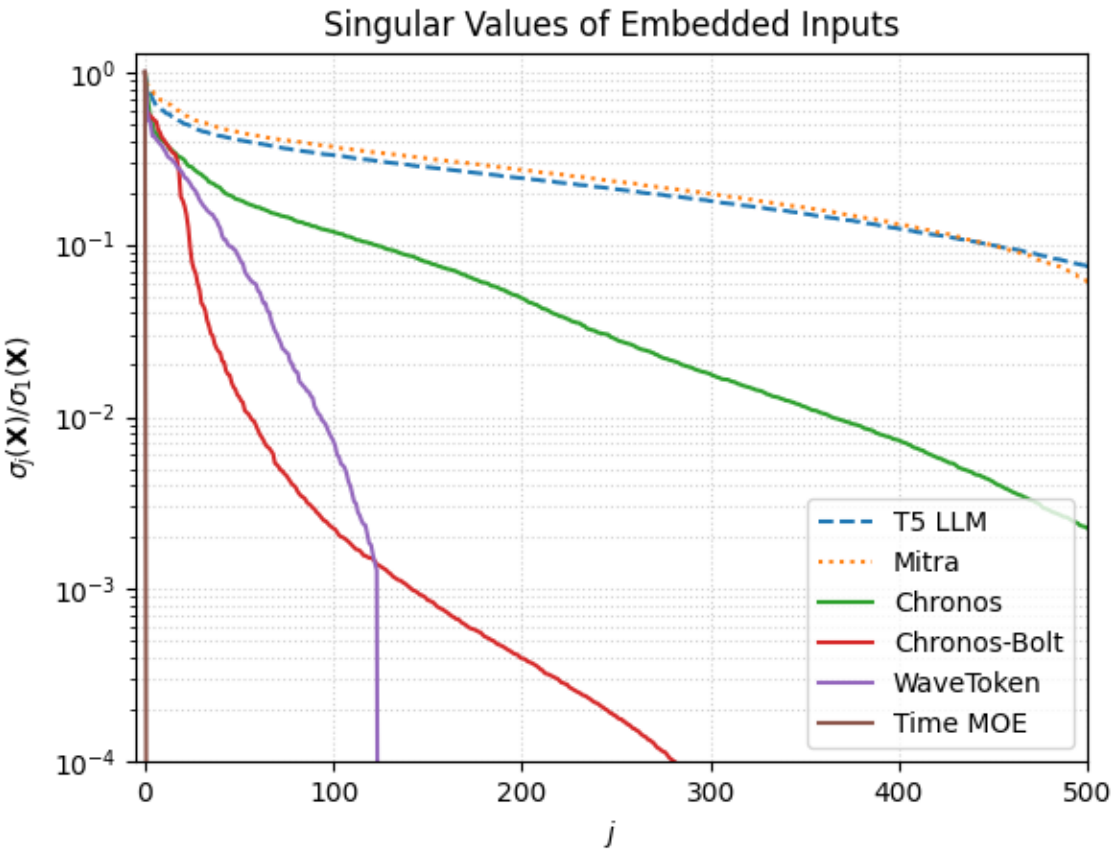
A Typical TSFM

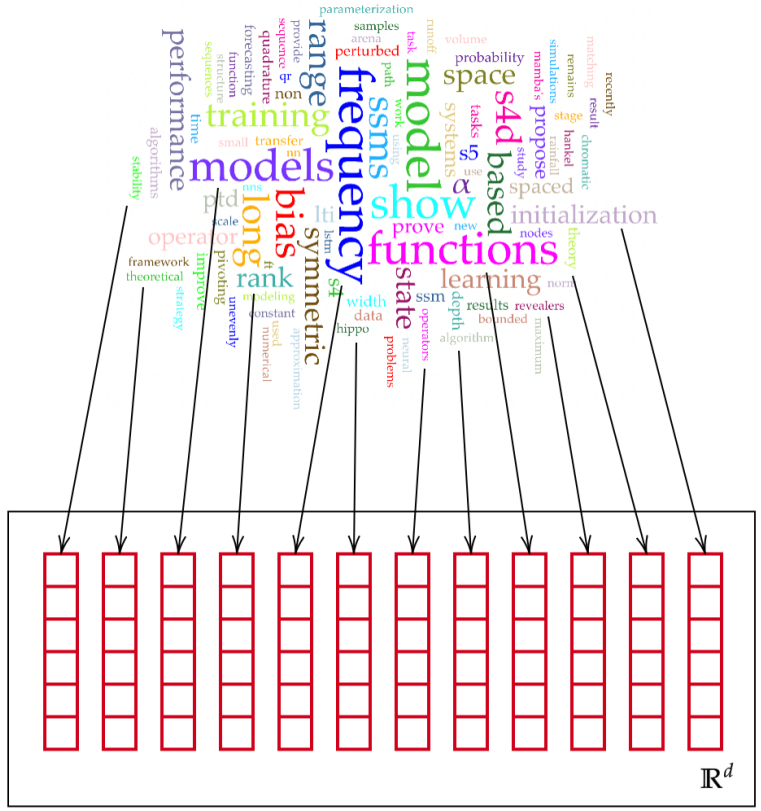


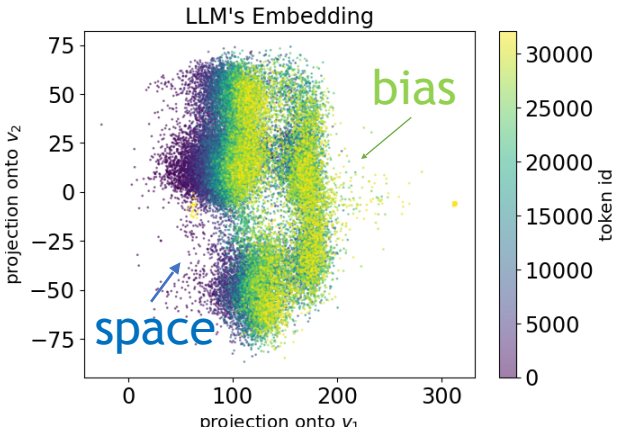
Low-rank Embeddings



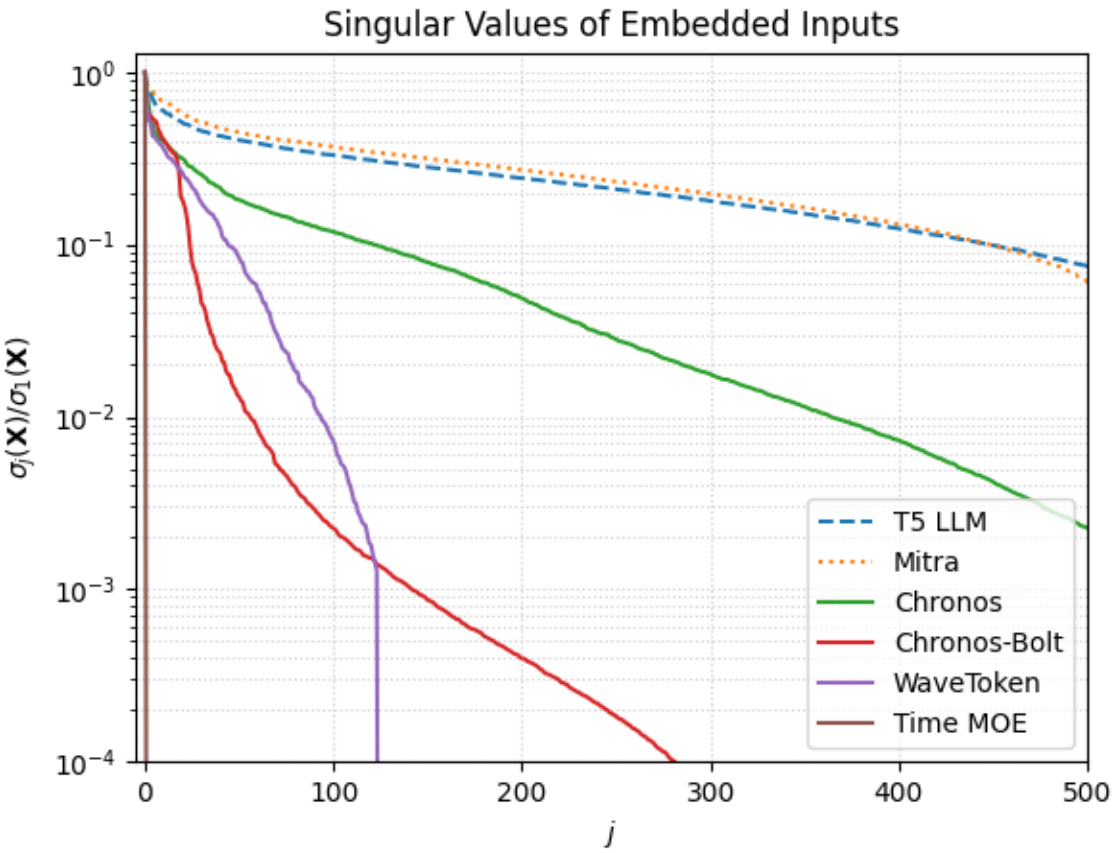
Low-rank Embeddings

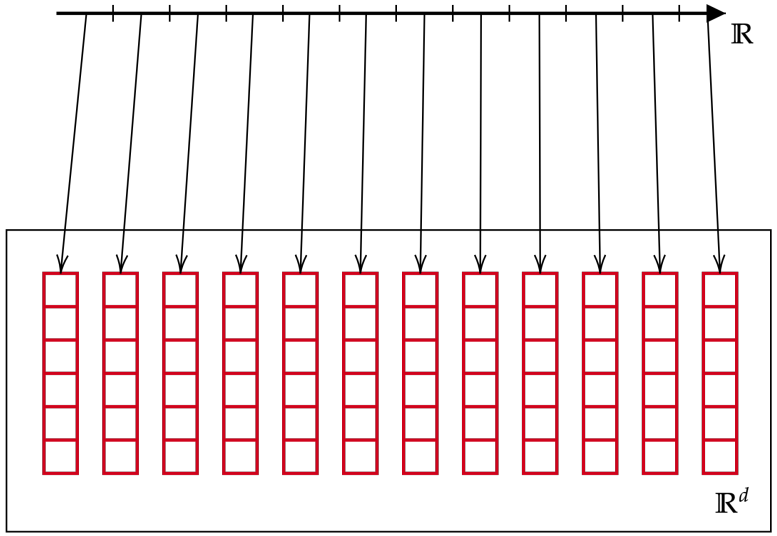


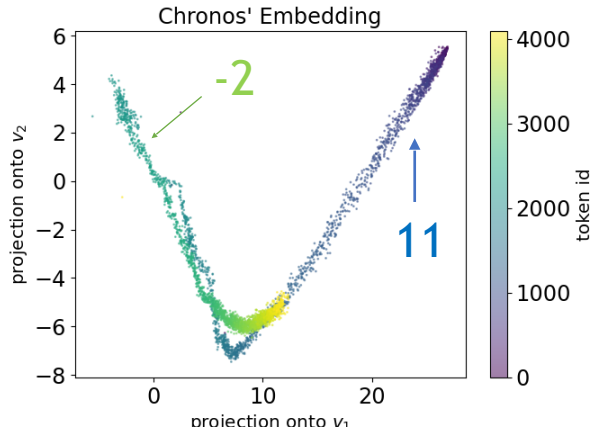




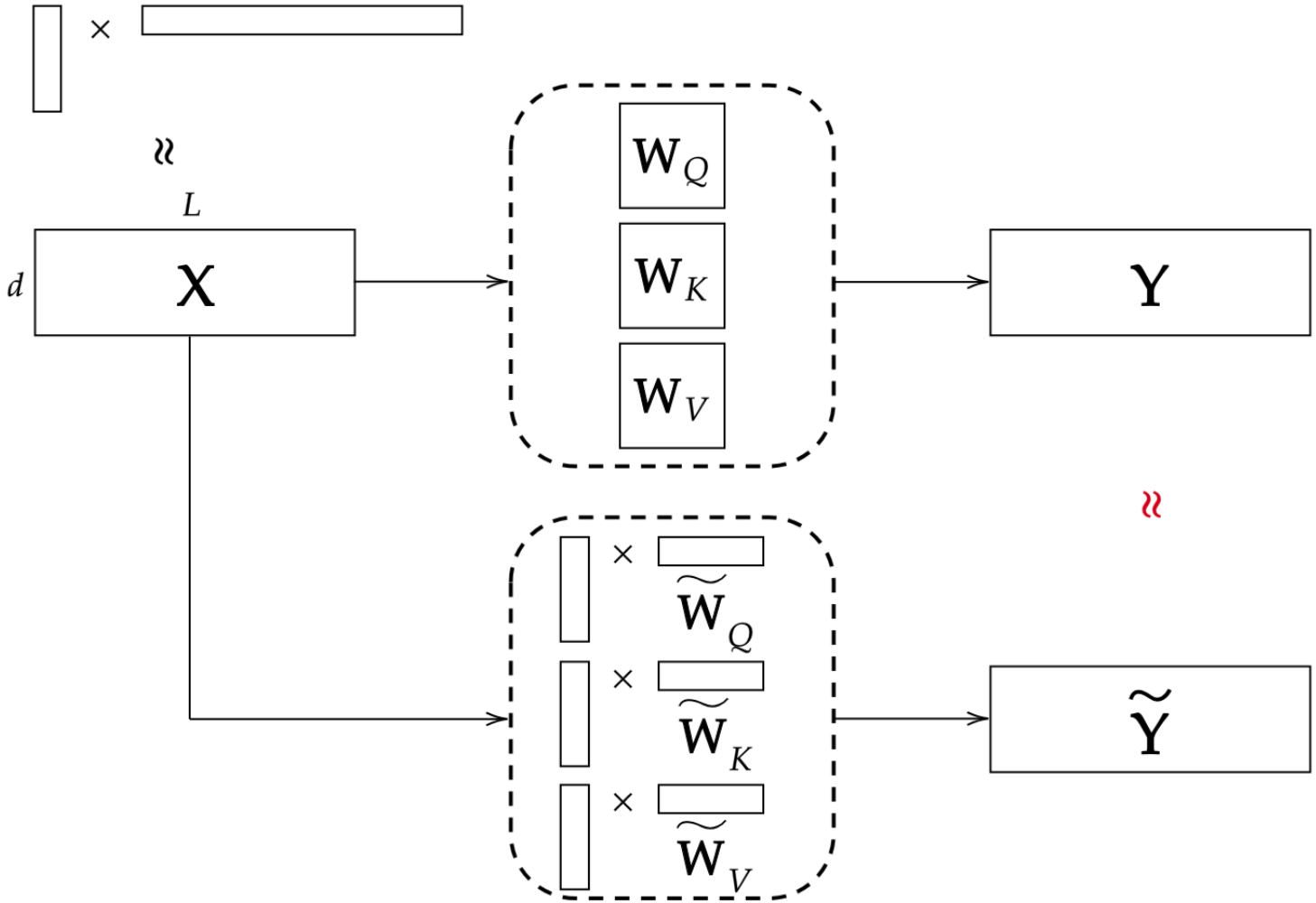
Low-rank Embeddings



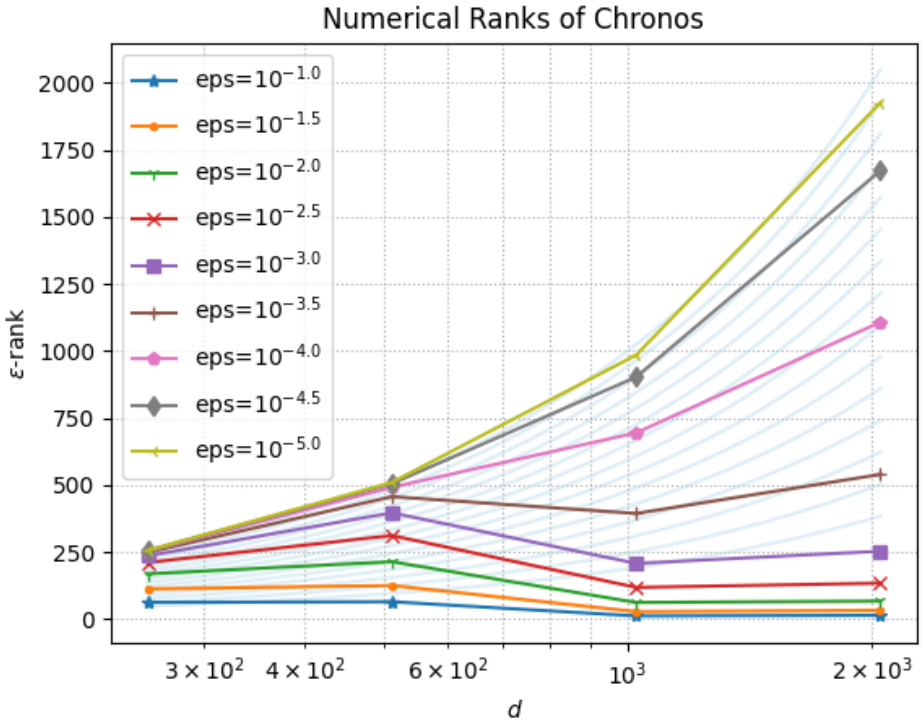


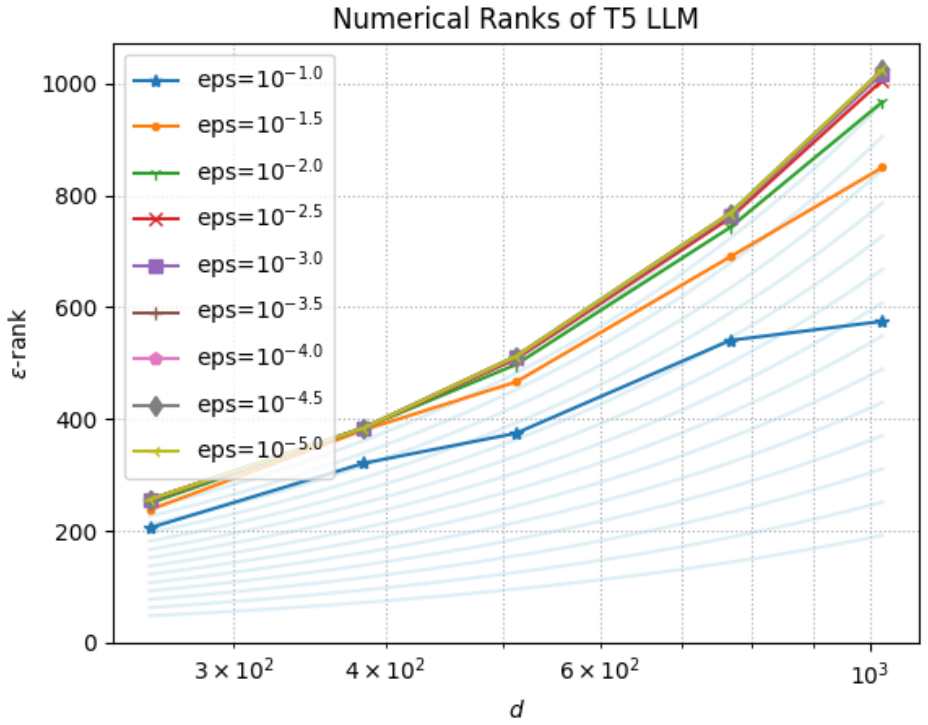


From Low-rank Inputs to Low-rank Attention



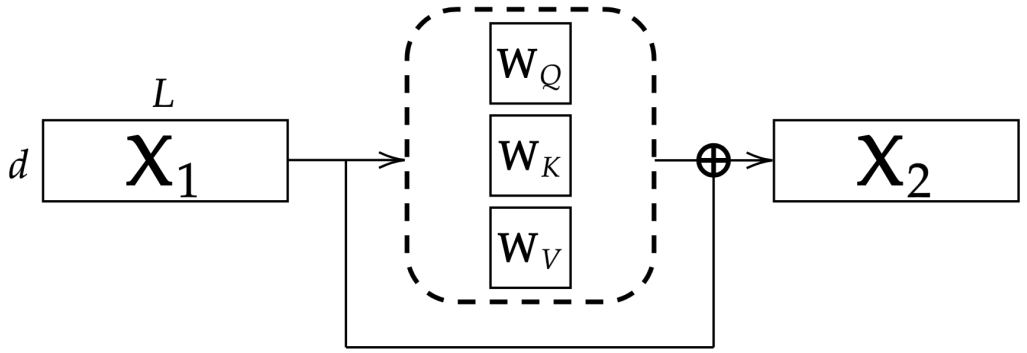
From Low-rank Inputs to Low-rank Attention



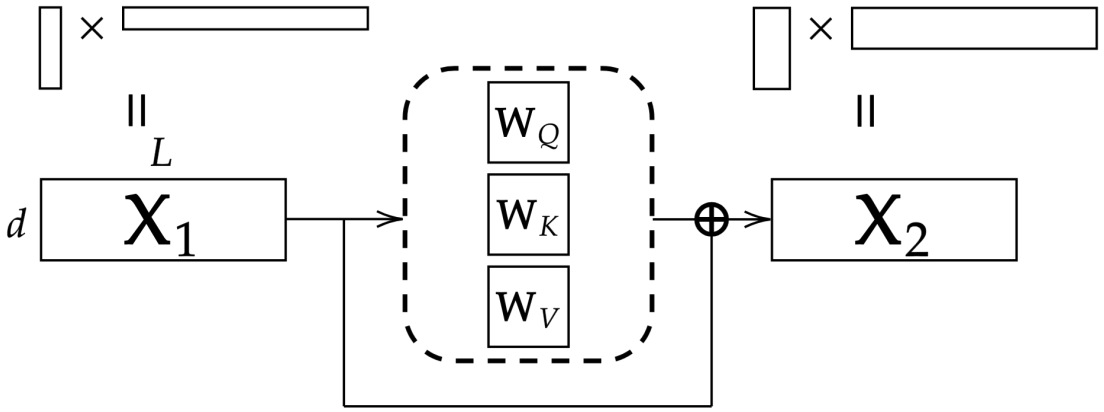


- TSFMs have low-rank attention matrices, but LLMs don’t.

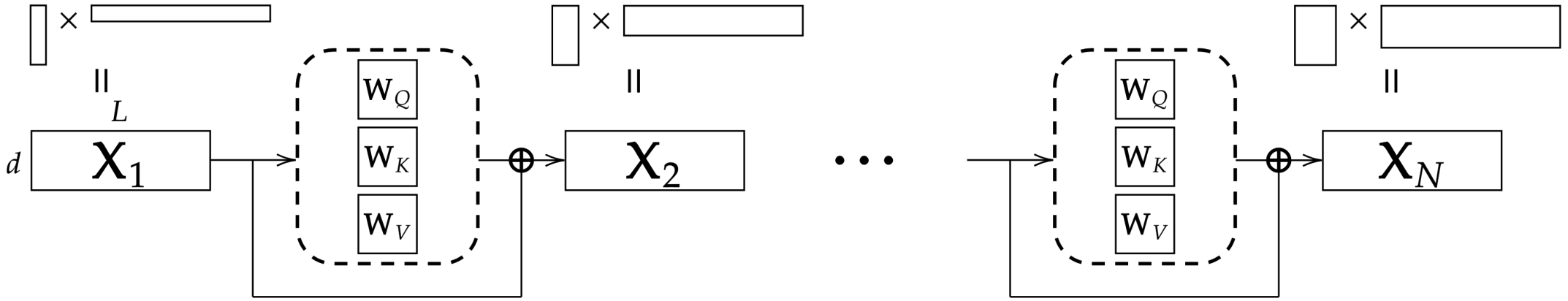
Flow of Ranks



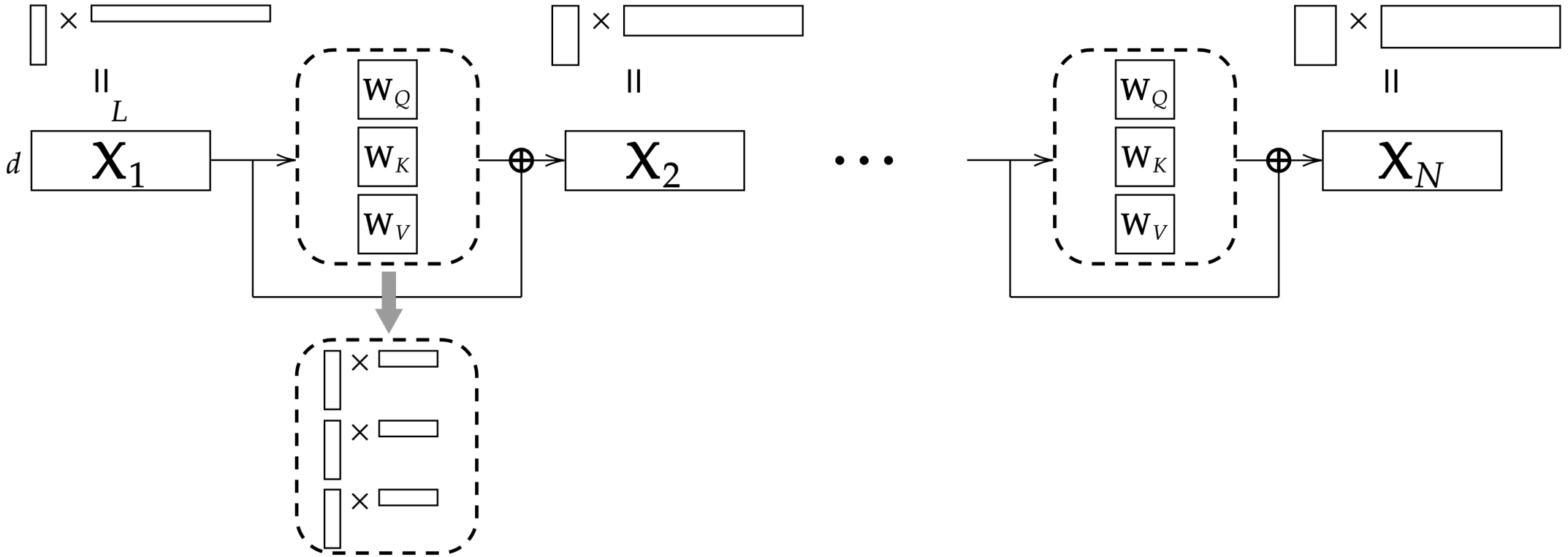
Flow of Ranks



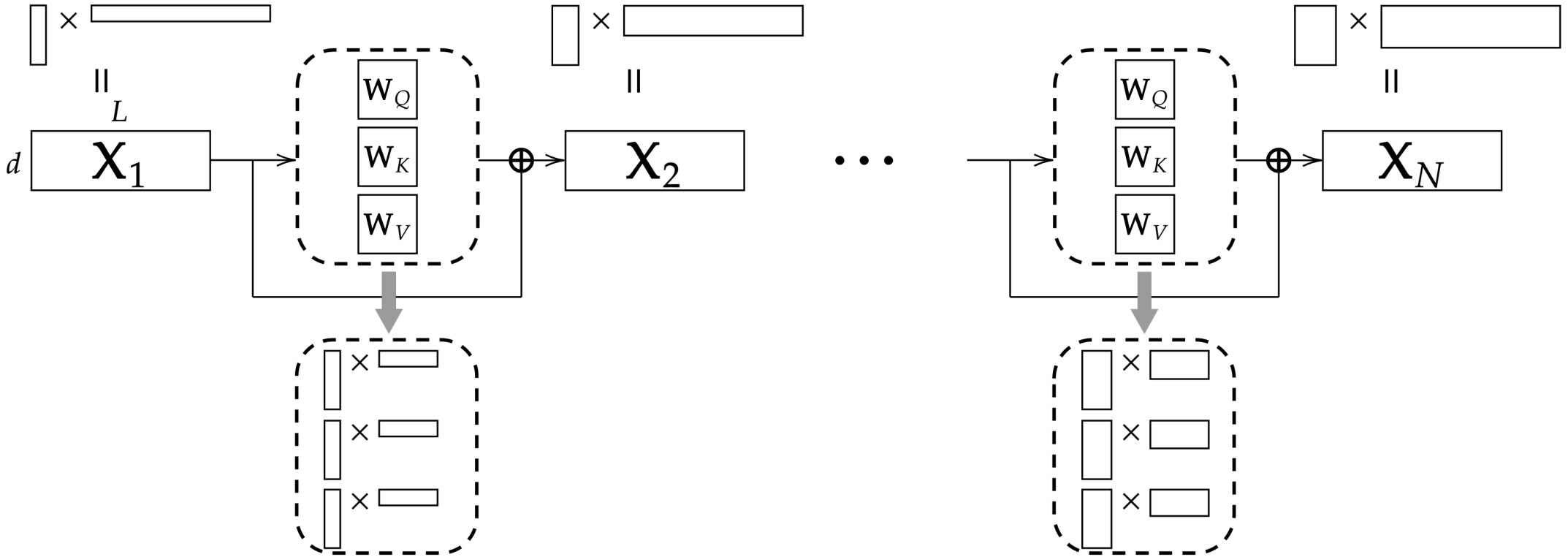
Flow of Ranks



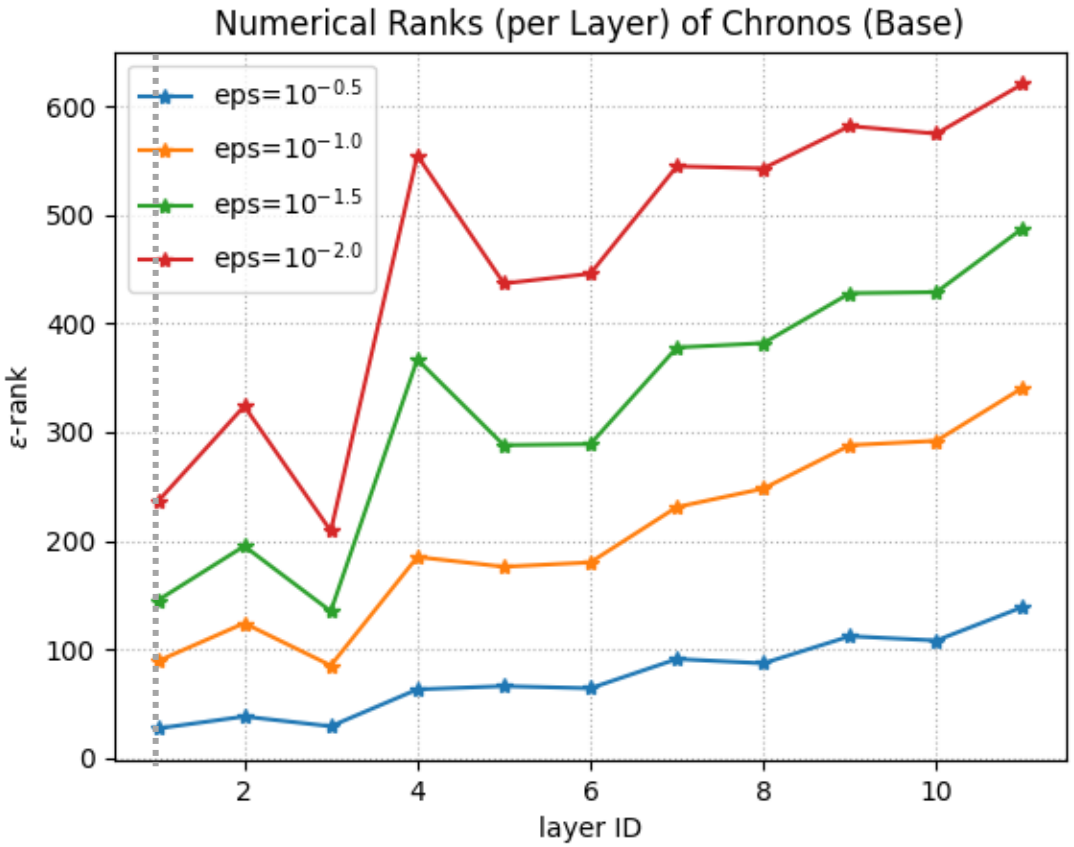
Flow of Ranks

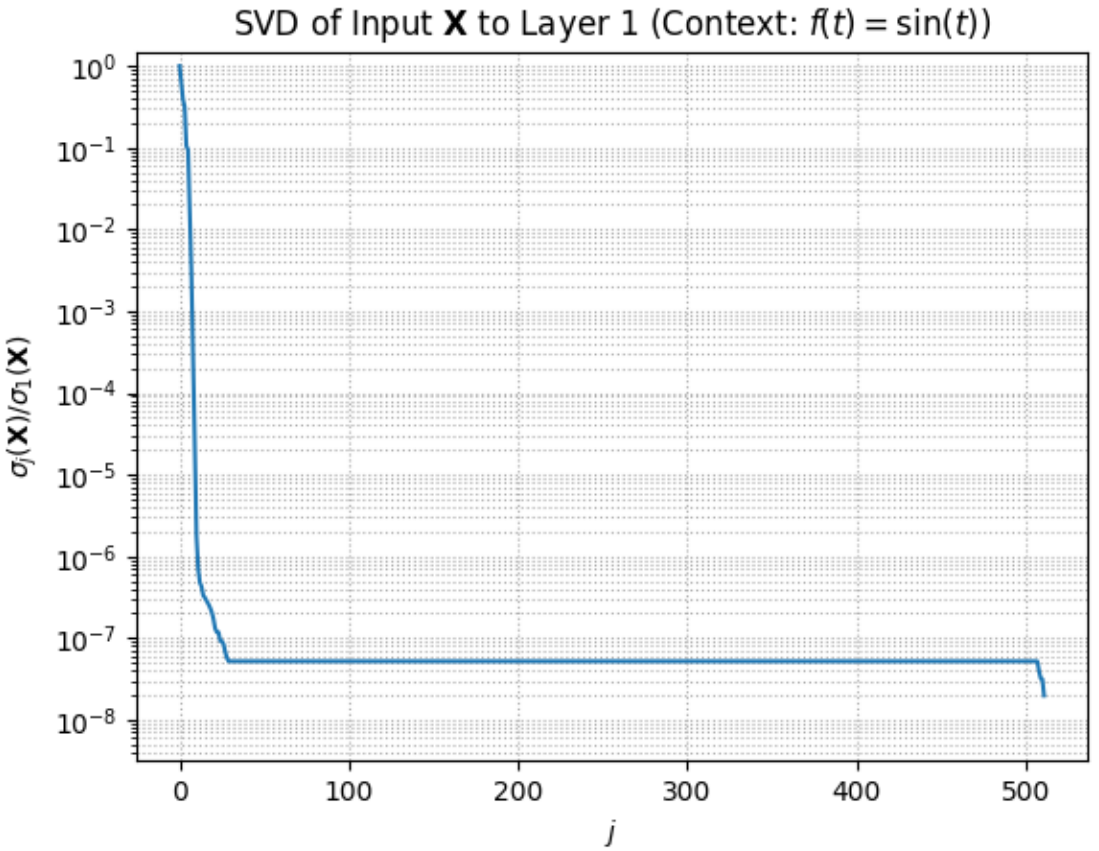


Flow of Ranks

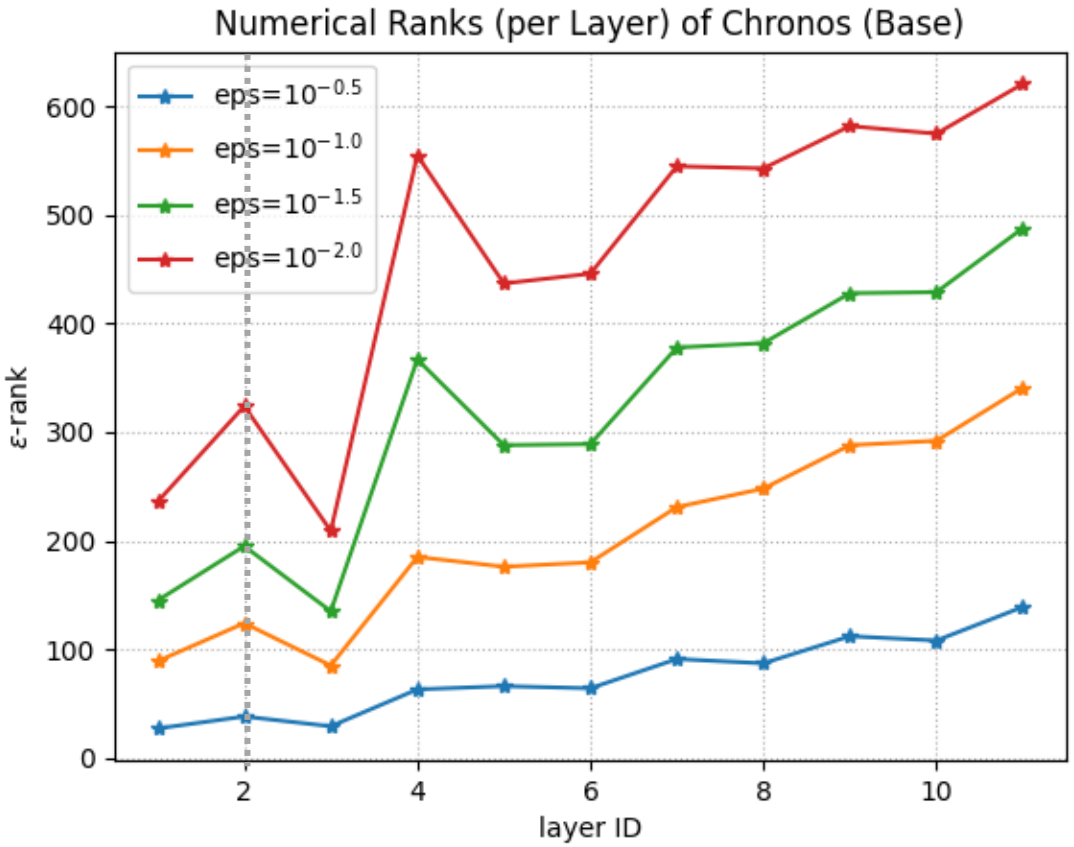


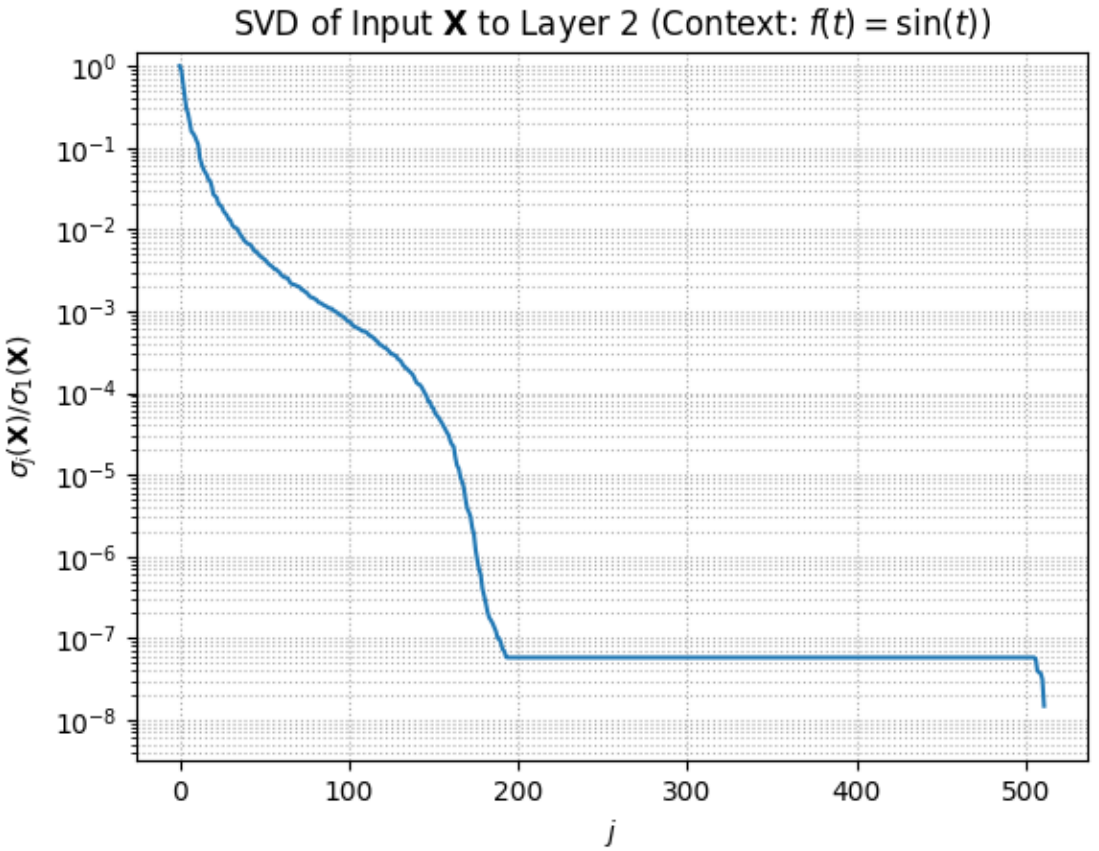
Flow of Ranks



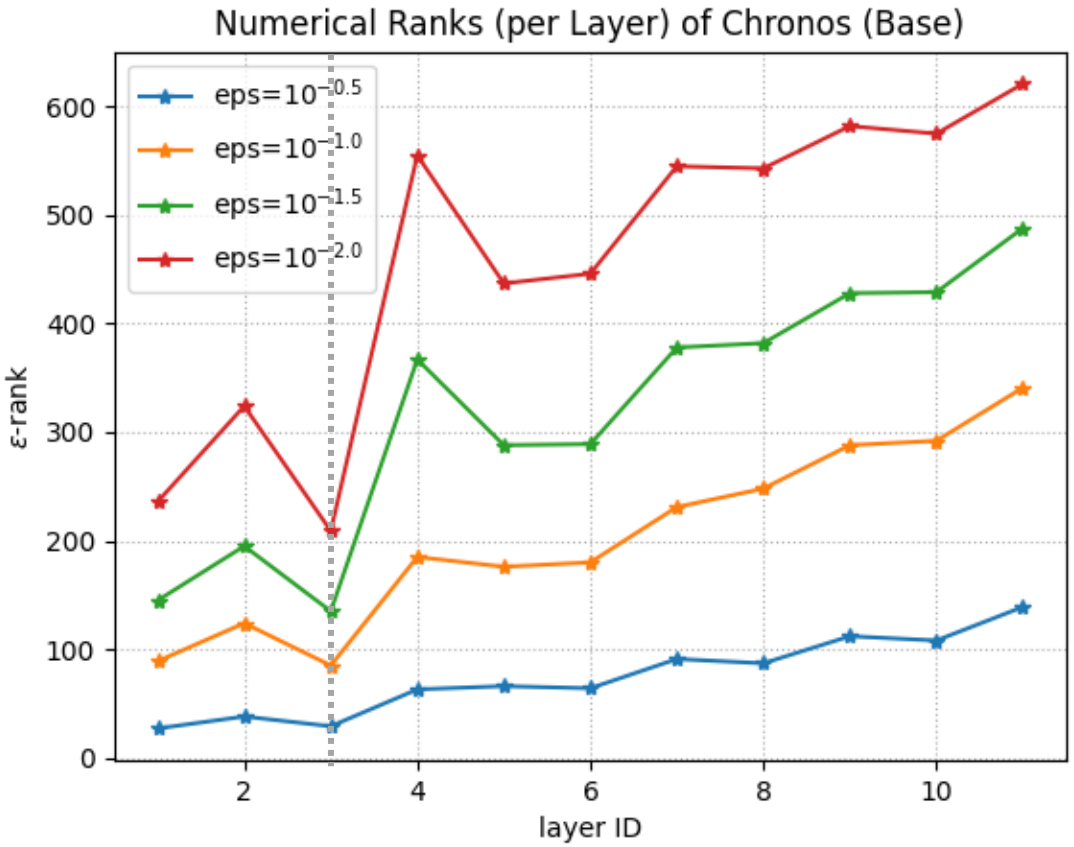


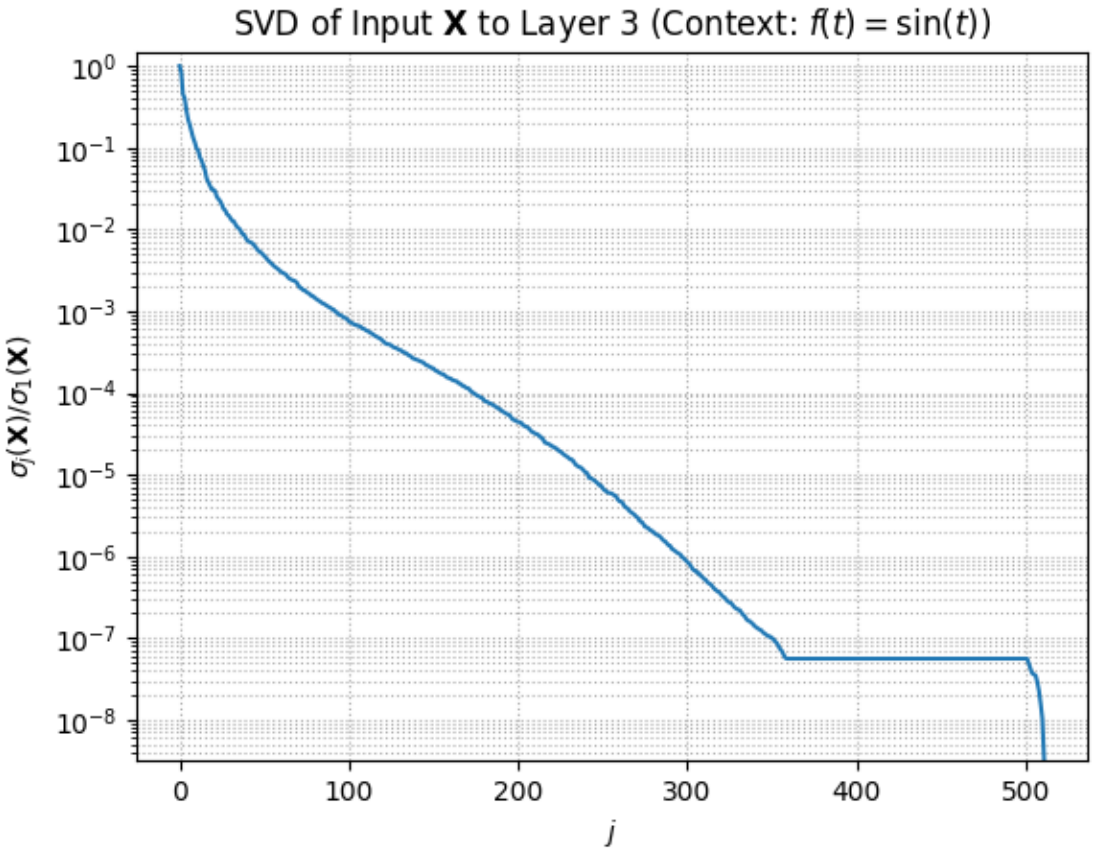
Flow of Ranks



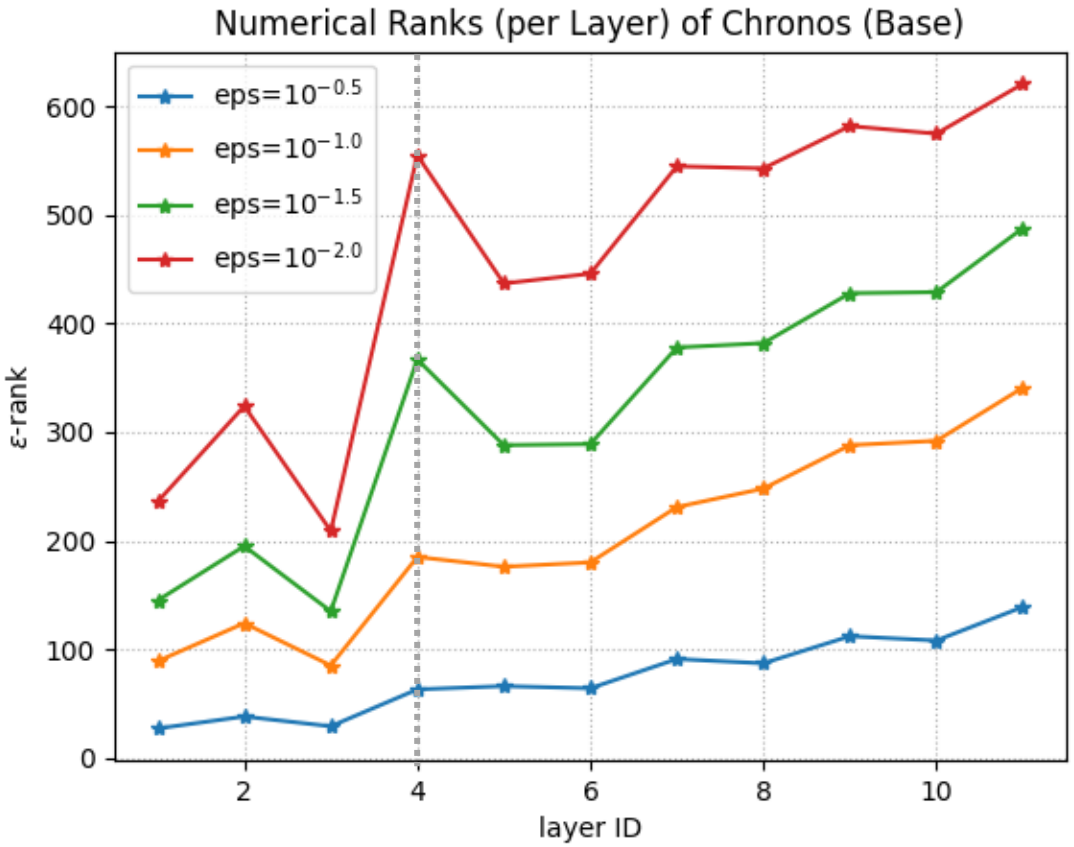


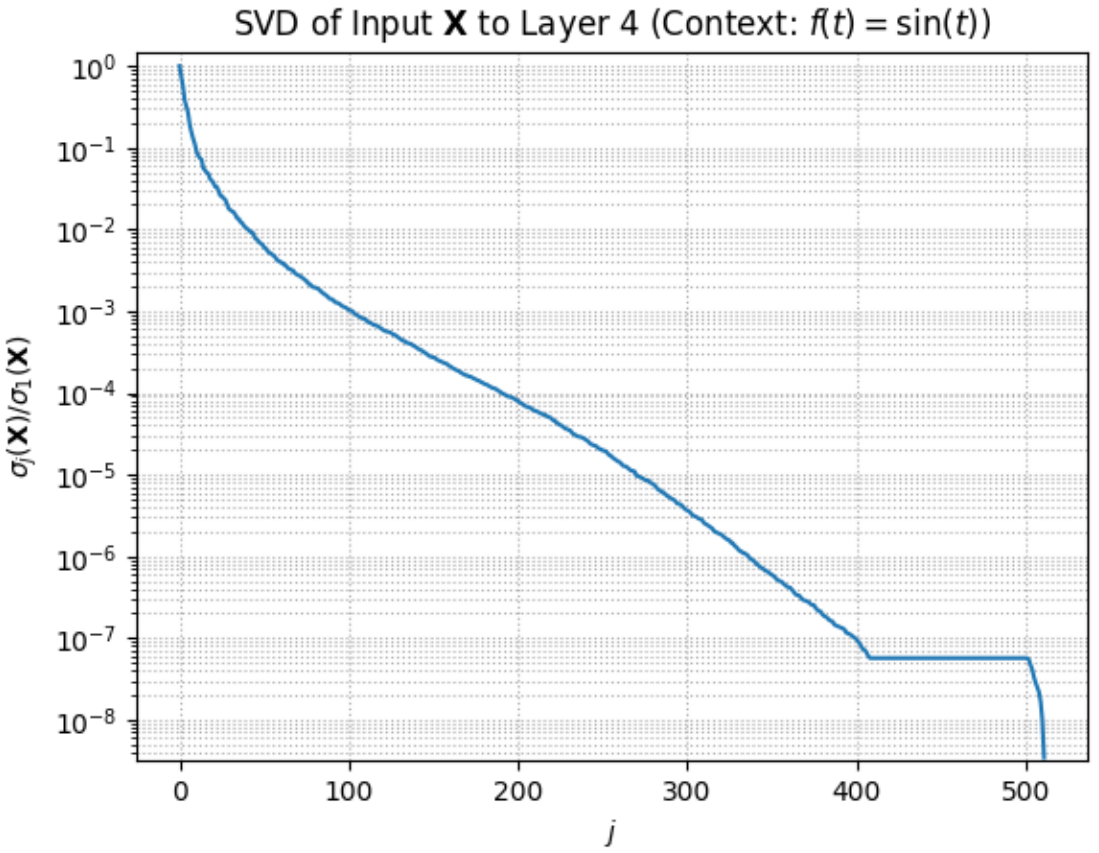
Flow of Ranks



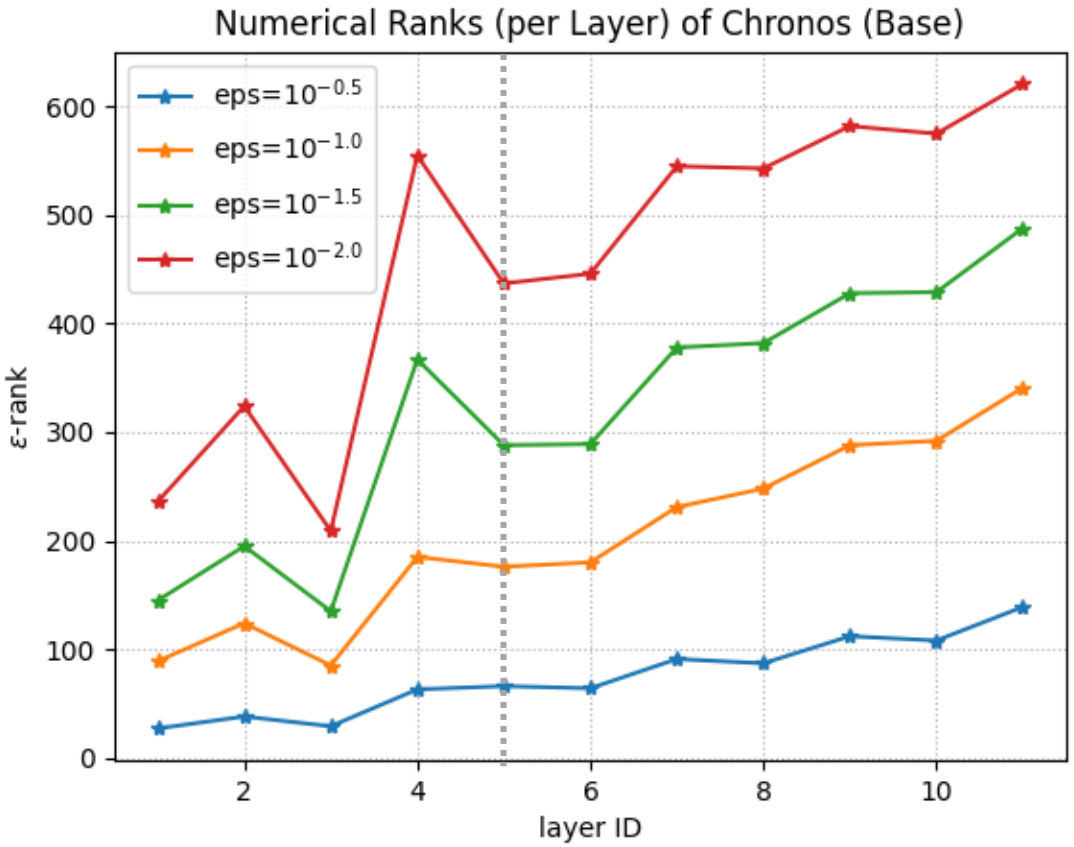


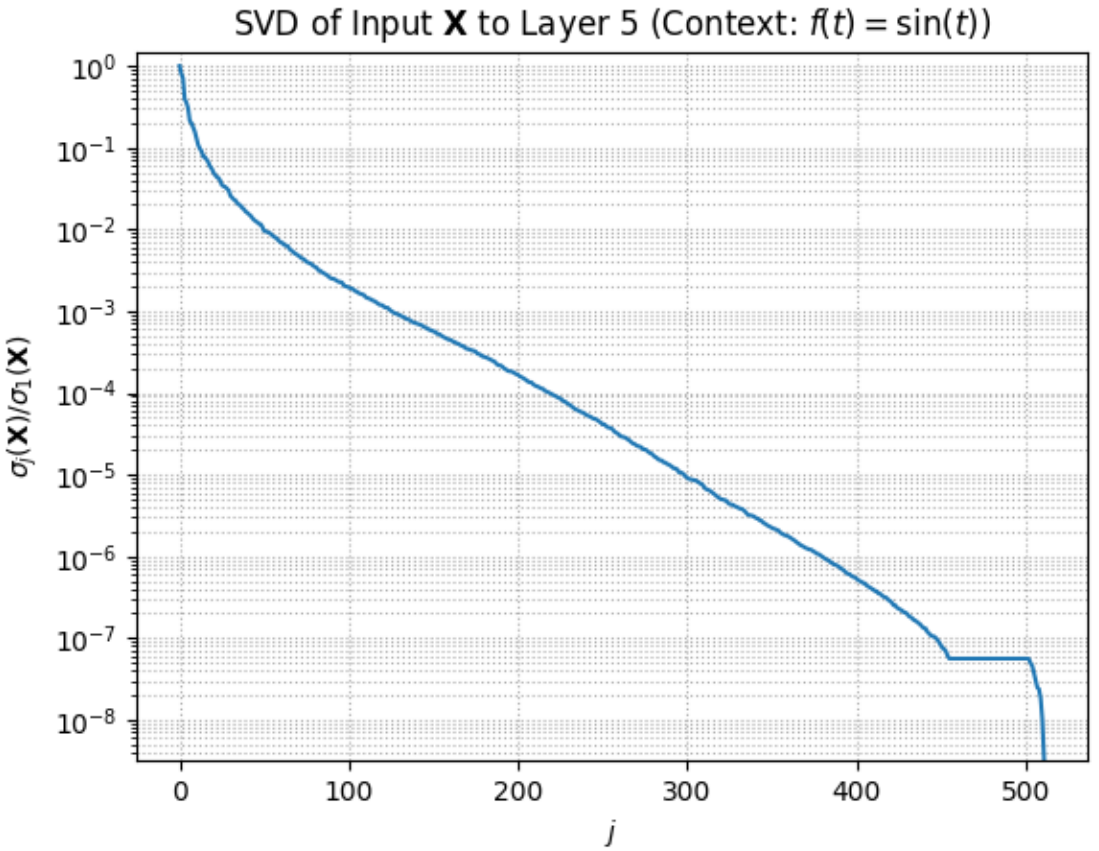
Flow of Ranks



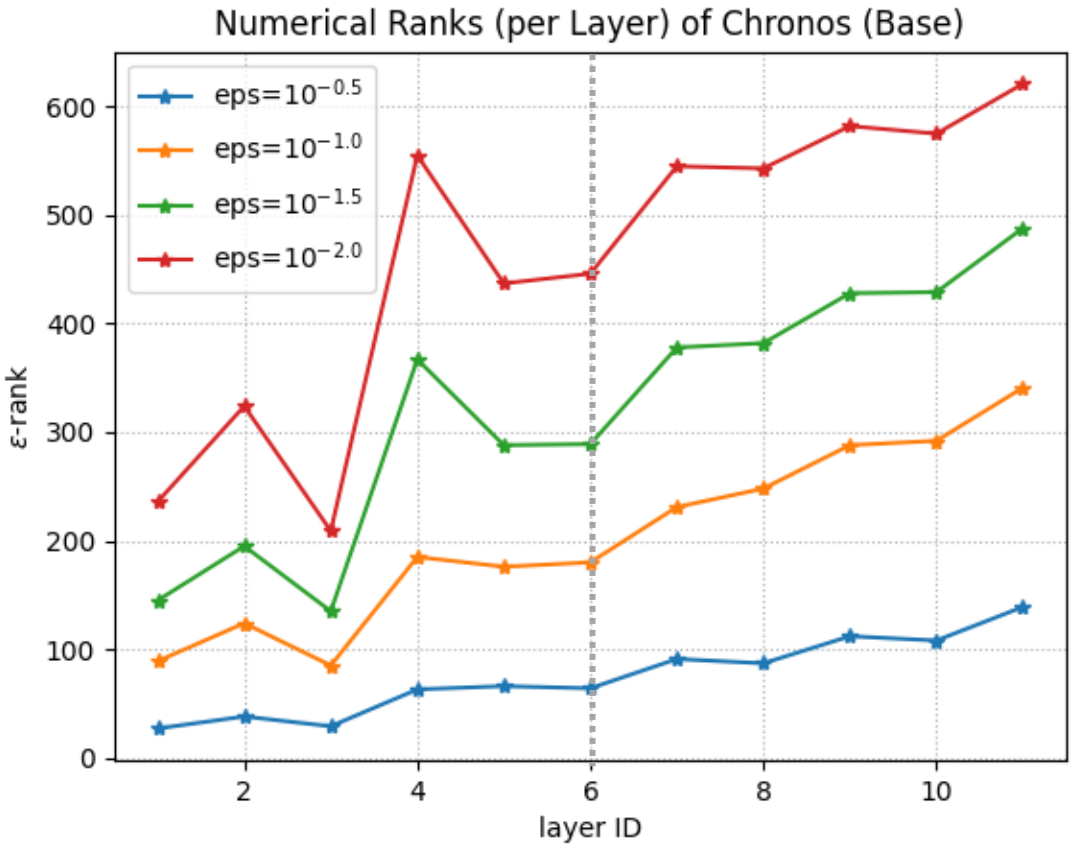


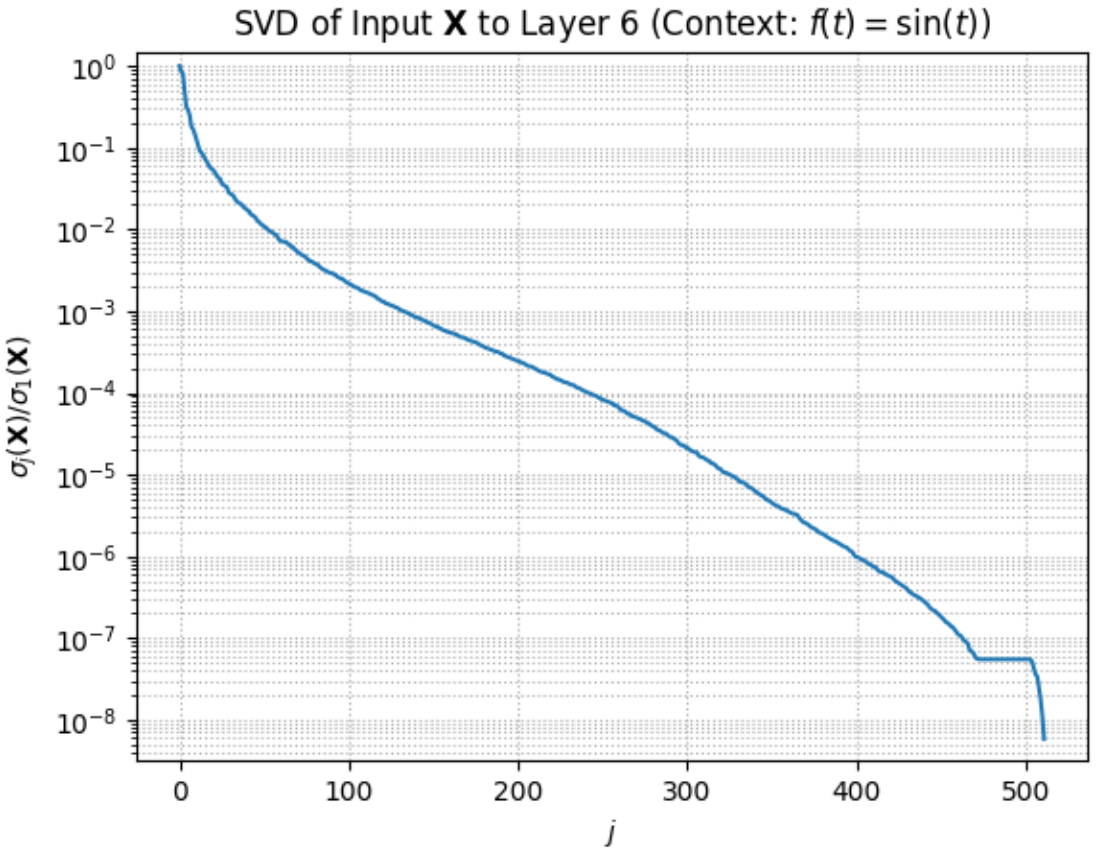
Flow of Ranks



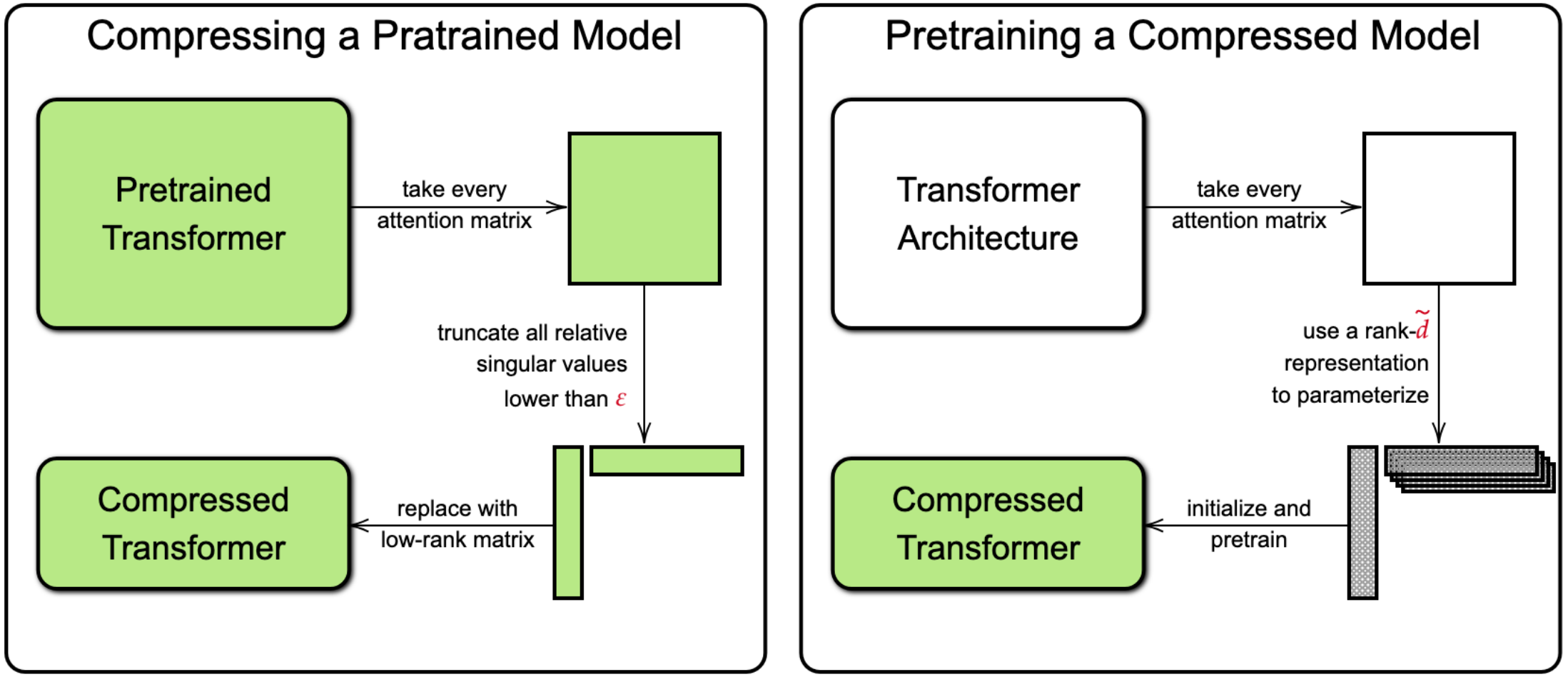


Flow of Ranks





Compressibility of TSFMs



Compressing Chronos reduces its inference time by 65% and its memory requirement by 81%, without loss of accuracy. More results (on Moirai, Chronos-Bolt, etc.) are found in our paper!