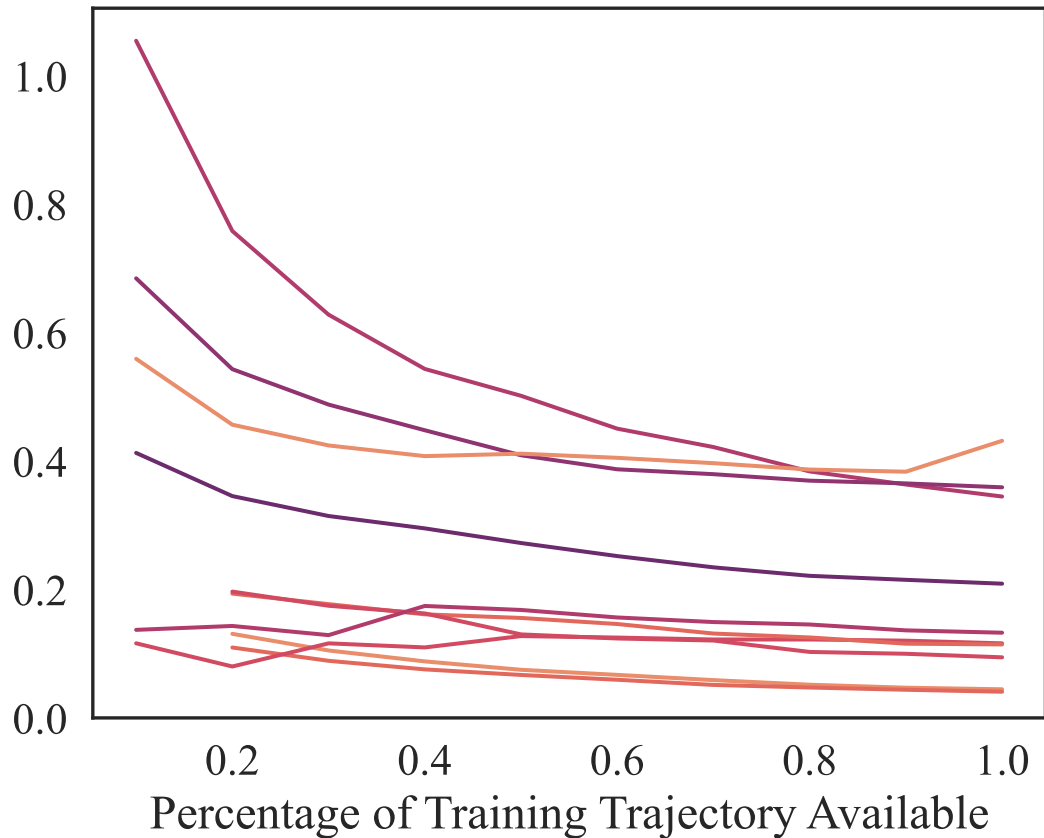


Mean Normalized Distance



data

- c4
- oscar
- MassiveText
- OPT(ROBERTA_PILE_REDDIT)
- pile-nodedup
- gpt3
- pile
- C4
- RedPajama
- RefinedWeb