Sequential Modelling of Typing Performance Logs

2. Dataset

The dataset used in this project has been collected from "**monkeytype**", an online typing platform that records detailed typing performance logs. For this project, two raw datasets were curated **User1** and **User2**. Together, these datasets capture typing behaviour across two different individuals, enabling both user specific and comparative analysis.

The dataset for User-1 contains 847 session records, while User-2 dataset contains 1000 records. Each record corresponds to a single typing session and consists of 24 attributes describing performance metrics, session characteristics, error statistics, and contextual metadata. The most central attributes include words per minute (**wpm**), raw words per minute (**rawWpm**), accuracy (**acc**), and consistency, which together quantify typing speed, correctness, and rhythm stability. In addition, the charStats field encodes detailed error information such as the number of correct, incorrect, extra, and missed characters.

Beyond performance, the dataset also contains information about the test environment and conditions. This includes the typing mode (mode and mode2), which specify the type and duration of the test, the length of the prompt (quoteLength), and the number of times a session was restarted (restartCount). Timing-related features such as testDuration, afkDuration (away from keyboard time), and incompleteTestSeconds provide insights into user behaviour during each session. Metadata such as language, difficulty, and timestamp allow the logs to be contextualised in terms of linguistic setting, difficulty level, and temporal trends.

It is worth noting that the dataset does contain missing values in some optional fields such as quoteLength, tags, and funbox, but the primary performance and timing metrics are consistently available. Overall, the dataset is rich in sequential log data, combining both quantitative performance indicators and contextual descriptors. This makes it well suited for exploratory data analysis, clustering of behavioural patterns, and predictive sequence modelling of typing performance trends.