All datasets come from the Hugging Face Datasets hub. For example, AG News is loaded in Python via from datasets import load_dataset and load_dataset("ag_news"); SST-2 via load_dataset("glue", "sst2"); and IMDB via load_dataset("imdb"). We pin revisions for reproducibility, use official splits (or a stratified 10% validation split when needed), do minimal cleaning (e.g., strip whitespace; remove
 tags). We keep these CSVs (plus checksums) locally for deterministic reruns, while sharing scripts and indices—not re-hosting the raw data