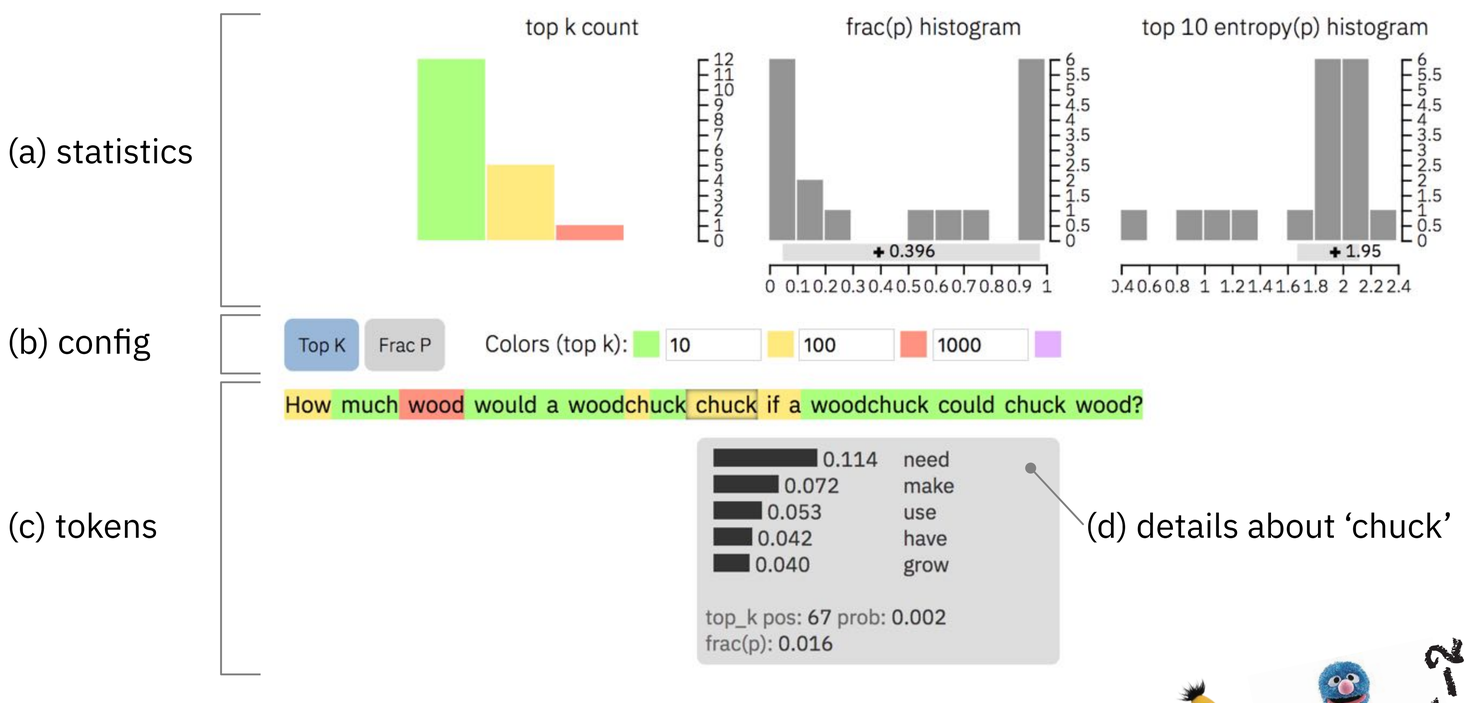
GLTR: Statistical Detection and Visualization of Generated Text

Sebastian Gehrmann*, Hendrik Strobelt* and Alexander Rush

The rapid improvement of language models has raised the specter of abuse of text generation systems. This progress motivates the development of simple methods for detecting generated text that can be used by non-experts. In this work, we introduce GLTR, a tool to support humans in detecting whether a text was generated by a model. GLTR applies a suite of baseline statistical methods that can detect generation artifacts across multiple sampling schemes. In a human-subjects study, we show that the annotation scheme provided by GLTR improves the human detection-rate of fake text from 54% to 72% without any prior training. GLTR is open-source and publicly deployed, and has already been widely used to detect generated outputs.

www.gltr.io

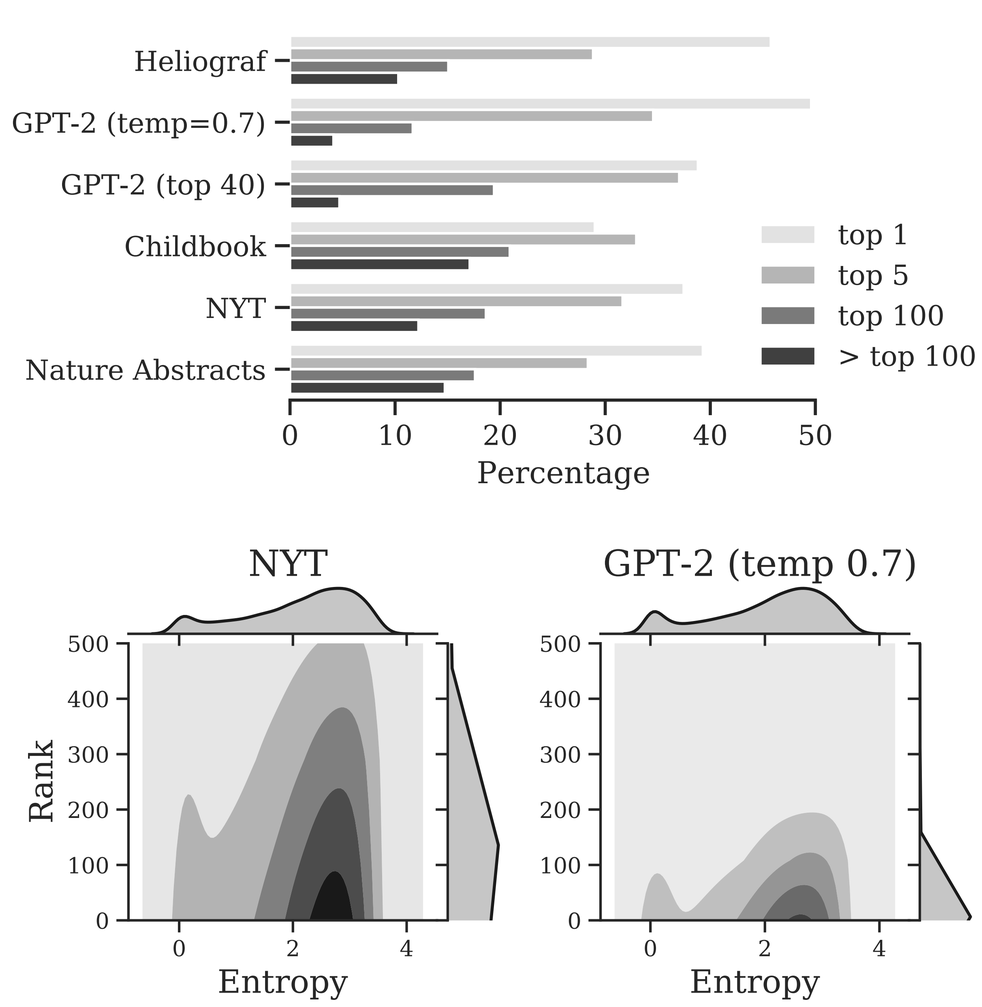
System



User interface for GLTR. On the top, we show three graphs with global information (a). Below thegraphs, users can switch between two different annotations and customize the top-k thresholds (b). On the bottom, each token is shown with the associated annotation as heatmap (c). The tooltip (d) highlights information aboutthe current prediction when hovering over the word “chuck”.

Example

Empirical Validation



MIT-IBM Watson AI Lab

HARVARD School of Engineering and Applied Sciences