CAREER: Towards Measuring and Improving Data Quality in Natural Language Processing

PI: Swabha Swayamdipta **Email:** swabhas@usc.edu
Affiliation: Assistant Professor, University of Southern California

---

Overview

[E.g., *Describe the activity that would result if the proposal were funded. State the objectives and methods to be employed*]

Keywords Natural langauge processing; Language Generation

Intellectual Merits

[E.g., *Potential to advance knowledge and understanding within and across fields • Qualifications of investigators • If collaboration, how chosen partners strengthen project • Creativity and originality • Conceptualization and organization • Access to resources*]

Broader Impacts

Our proposal could lead to the design of practical data-efficient algorithms for training large language models. We will develop open-source software implementations. Alongside pursuing such broader impacts in research, we will pursue several educational and outreach activities which are tied together with our research goals. Two key highlights are (1) a partnership with Los Angeles County Office of Education and Code.org to develop high-school lessons on the role of language models, (2) mentoring local high-school students and visiting undergraduates over the summer on research topics motivated by the proposal. Across these activities we will especially aim to reach groups which are historically underrepresented in computer science.