

# Semantic Token Clustering for Efficient Uncertainty Quantification in Large Language Models

東京大学大学院 工学系研究科 氏名 Qi Cao

目的 The paper aims to develop an efficient, self-contained method for quantifying LLM output uncertainty from a single generation without auxiliary models or repeated sampling.

内容 It proposes Semantic Token Clustering (STC), an efficient uncertainty quantification method that leverages the semantic information inherently encoded in LLMs. STC requires only a single generation and does not depend on auxiliary models.

結果 Experimental results show that STC achieves performance comparable to state-of-the-art baselines while substantially reducing computational overhead.

利用した計算機 SQUID 汎用 CPU ノード群  
SQUID GPU ノード群

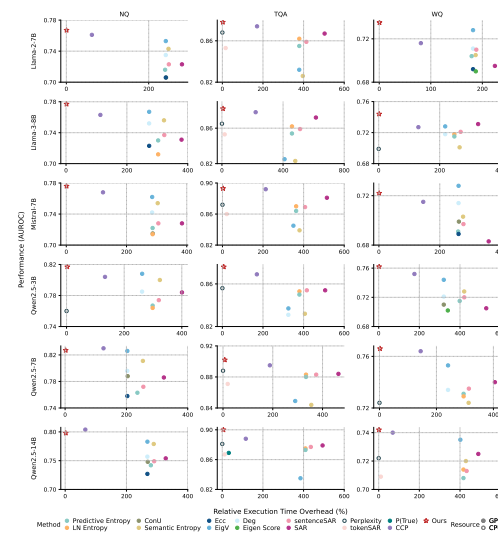


Figure. Efficiency comparison.