

Detecting LLM Hallucinations via Semantic Entropy

Luka Aghapishvili^a

e-mail: luka.aghapishvili897@ens.tsu.edu.ge

^a Computer Science, Faculty of Exact and Natural Sciences,
Tbilisi State University, Ilia Chavchavadze 1s

While modern Large Language Models (LLMs) generate complex text with high linguistic confidence, they frequently present semantically incorrect or hallucinated information despite maintaining a plausible form. The primary objective of this project is to identify LLM hallucinations by utilizing Semantic Entropy, a method that enables the measurement of meaning divergence across varied responses.

The methodology is based on multiple stochastic generations, the conversion of responses into semantic vectors (embeddings), clustering by meaning, and the calculation of entropy based on Information Theory. The resulting Normalized Semantic Entropy (NSE) facilitates the detection of hallucinations, enhances identification accuracy, and ensures contextual consistency. The project includes an MVP version that integrates LLM APIs, embeddings, and clustering algorithms to demonstrate real-world semantic diversification. The expected outcome is a relevant, adaptive, and robust system suitable for both research and practical applications.