

სემანტიკურ ენტროპიაზე დაფუძნებული LLM ჰალუცინაციების იდენტიფიცირება

ლუკა ალაბიშვილი^ა

ელ-ფოსტა: luka.aghapishvili897@ens.tsu.edu.ge

^ა კომპიუტერული მეცნიერება, ზუსტი და საბუნებისმეტყველო ფაკულტეტი, თბილისის სახელმწიფო უნივერსიტეტი, ილია ჭავჭავაძის 1

თანამედროვე დიდი ენობრივი მოდელები (LLM) მაღალი დაჯერებულობით აგენერირებენ რთულ ტექსტებს, თუმცა, მიუხედავად სარწმუნო ფორმისა, ისინი ხშირად წარმოადგენენ სემანტიკურად მცდარ ან ჰალუცინირებულ ინფორმაციას. წინამდებარე პროექტის მთავარი მიზანია LLM ჰალუცინაციების იდენტიფიცირება სემანტიკური ენტროპიის გამოყენებით, რაც საშუალებას გვაძლევს გავზომოთ მნიშვნელობების დივერგენცია მოდელის მიერ გენერირებულ სხვადასხვა პასუხში.

აღნიშნული მეთოდოლოგია ეფუძნება მრავალჯერად სტოკასტურ გენერაციას, პასუხების კონვერტირებას სემანტიკურ ვექტორებად, მათ კლასტერიზაციასა და ინფორმაციულ თეორიაზე დაფუძნებული ენტროპიის გამოთვლას. მიღებული ნორმალიზებული სემანტიკური ენტროპია (NSE) ეფექტურად ავლენს ჰალუცინაციების რისკს, ზრდის დეტექციის სიზუსტეს და უზრუნველყოფს კონტექსტურ თანმიმდევრულობას. პროექტის ფარგლებში შემუშავებული MVP ვერსია აერთიანებს LLM API-ს, ემბედინგებსა და კლასტერიზაციის ალგორითმებს რეალური სემანტიკური დივერსიფიკაციის წარმოსაჩენად. მოსალოდნელი შედეგია ადაპტირებადი და მაღალტექნოლოგიური სისტემა, რომელიც წარმატებით იქნება გამოყენებული როგორც სამეცნიერო კვლევებში, ისე პრაქტიკულ აპლიკაციებში.