Modelo de linguagem LLama

Cristiano de Almeida Tomaz

Instituto Federal de Educação Ciência e Tecnologia de São Paulo]
Campos do Jordão – SP – Brasil

Departamento de Análise e Desenvolvimento de Sistemas

cristiano.tomaz@ifsp.edu.br

Abstract. *Artificial intelligence-based language models have had a significant impact on society in recent years. Their task automation and information generation capabilities are increasingly impressive due to their growing speed and accuracy. In this context, LLama, a free-access language model developed by Meta AI, stands out. LLama has been designed to provide competitive performance in text comprehension and generation tasks, using an efficient and accessible architecture. It emerges as an open and efficient alternative, trained exclusively on publicly available data. With an optimized architecture and scaled between 7 billion and 65 billion parameters, LLama seeks to strike a balance between high performance and accessibility, allowing it to run on less robust hardware, thus democratizing access to research with advanced language models. In this article, we will explore the workings of the LLama model, its applications, and impacts in the field of systems development. The methodologies used in the construction of the model, its advantages over other language models, and its limitations will be discussed, as well as the potential use of this technology in developing customized solutions.*

Resumo. *Os modelos de linguagem baseados em inteligência artificial têm gerado grande impacto na sociedade nos últimos anos. Seus recursos de automatização de tarefas e geração de informação impressionam pela velocidade e precisão cada vez maiores. Neste cenário tem destaque o LLama, um modelo de linguagem de acesso gratuito desenvolvido pela Meta AI. O LLama tem sido projetado para fornecer um desempenho competitivo em tarefas de compreensão e geração de texto, utilizando uma arquitetura eficiente e acessível. E surge como uma alternativa eficiente e aberta, treinada exclusivamente em dados disponíveis publicamente. Projetado com uma arquitetura otimizada e dimensionado entre 7 bilhões e 65 bilhões de parâmetros, o LLaMA busca alcançar um equilíbrio entre alto desempenho e acessibilidade, permitindo sua execução em hardware menos robusto. E assim democratizando o acesso à pesquisa com modelos de linguagem avançados. Neste artigo vamos explorar o funcionamento do modelo LLama, suas aplicações e impactos no campo do desenvolvimento de sistemas. Serão abordadas as metodologias empregadas na construção do modelo, as vantagens em relação a outros modelos de linguagem e suas limitações, além de discutir o potencial uso desse tipo de tecnologia no contexto no desenvolvimento de soluções personalizadas.*