

Perancangan MLOps untuk model Generative AI dengan pengujian berbasis Chaos Engineering = MLOps design for Generative AI models with Chaos Engineering-based testing

Bisma Khomeini, author

Deskripsi Lengkap: <https://lib.ui.ac.id/detail?id=9999920552886&lokasi=lokal>

Abstrak

Pembelajaran mesin telah mengubah berbagai industri dan aplikasi secara mendasar, membuka era baru dalam pemanfaatan data. Kemampuannya dalam mengenali pola dan menganalisis informasi dari data besar telah menciptakan revolusi dalam berbagai bidang. Namun, seiring dengan pertumbuhan cepat dalam ranah machine learning, model-model yang semakin berkembang menjadi lebih besar dan kompleks, yang pada akhirnya meningkatkan tantangan dalam proses pelatihan dan manajemennya. Hal ini menghadirkan tantangan tersendiri dalam implementasi operasi pembelajaran mesin dalam skala besar. Untuk memenuhi Service Level Objective (SLO) sebuah perusahaan, diperlukan algoritma penjadwalan (scheduling algorithm) yang mampu mempercepat proses tersebut. Selain itu, untuk mengukur efektivitas algoritma scheduling, diperlukan pengujian yang komprehensif. Dalam konteks ini, teknik chaos engineering mampu menguji seberapa baik sistem yang telah dibangun beserta algoritma penjadwalan yang diterapkan.

.....Machine learning has fundamentally transformed various industries and applications, ushering in a new era in data utilization. Its ability to recognize patterns and analyze information from large datasets has revolutionized numerous fields. However, with the rapid growth in the realm of machine learning, increasingly complex and larger models pose challenges in their training and management processes. This presents unique challenges in implementing MLOps on a large scale. To meet a company's Service Level Objective (SLO), a scheduling algorithm capable of expediting these processes is required. Moreover, comprehensive testing is necessary to measure the effectiveness of the scheduling algorithm. In this context, chaos engineering can assess the system's efficacy alongside the applied scheduling algorithm.