Abstrak

Perbaikan kode otomatis adalah tugas penting dalam pengembangan perangkat lunak untuk mengurangi *bug* secara efisien. Penelitian ini berfokus pada pengembangan dan evaluasi teknik Chain-of-Thought (CoT) Prompting untuk meningkatkan kemampuan Large Language Models (LLM) dalam tugas Automated Program Repair (APR). CoT Prompting adalah teknik yang memandu LLM untuk menghasilkan penjelasan langkah demi langkah sebelum memberikan jawaban akhir, sehingga diharapkan dapat meningkatkan akurasi dan kualitas perbaikan kode. Penelitian ini menggunakan dataset QuixBugs untuk mengevaluasi performa beberapa model LLM, termasuk DeepSeek-V3 dan GPT-4o, dengan dua metode prompting, yaitu Standard Prompting dan CoT Prompting. Evaluasi dilakukan berdasarkan rata-rata jumlah *plausible patches* yang dihasilkan serta estimasi biaya penggunaan token. Hasil menunjukkan bahwa CoT Prompting meningkatkan performa pada sebagian besar model. DeepSeek-V3 mencatat performa tertinggi dengan rata-rata 36,6 *plausible patches* dan biaya terendah sebesar $0,006. GPT-4o juga menunjukkan hasil yang kompetitif dengan rata-rata 35,8 *plausible patches* dan biaya $0,226. Hasil ini menegaskan bahwa CoT Prompting adalah teknik yang efektif untuk meningkatkan kemampuan reasoning LLM dalam tugas APR.

Kata kunci: chain-of-thought prompting, automated program repair, large language models, quixbugs