# *ABSTRACT*

*The field of Natural Language Processing (NLP) has experienced remarkable advancements, particularly in task Question Answering (QA), enhancing user interaction with systems. However, a significant challenge persists in the development of QA systems: ensuring that a system provides accurate answers while considering context. Specifically, Indonesian QA system, categorized as a low-resource language, encounters limitations in available datasets. This affects the model's performance, especially when confronted with language variations and complexities not fully represented in the dataset.*

*One solution to address these limitations is translating dataset from resource-rich languages (e.g., English) to Indonesian. The utilization of Transformer models, such as T5, in QA system development has been a focal point of related research. T5 allows for the implementation of transfer learning without significant architectural adjustments, and the text-to-text format provides considerable flexibility in adapting the model to various NLP domains and tasks, including QA. Unfortunately, a primary constraint of the T5 model is its demand for substantial computational resources, given its parameter count of 246,078,720. This poses a challenge, especially for users lacking access to high-level computing or sufficient infrastructure. Therefore, this study aims to compress the number of T5 model parameters to make it viable for limited computational resources.*

*Based on experimental results, the integration of Q-LoRA during the fine-tuning process with the T5 model successfully reduced the parameter count to 1,769,472, significantly lower than the original T5 model. Furthermore, the integration of Q-LoRA also yielded an improvement in inference time efficiency. This step not only reduced memory requirements but also enhanced computational efficiency, enabling the utilization of resources with limited computing capabilities.*

***Keywords:*** *Natural Language Processing, Question Answering, Compression Method, Text-to-Text Transfer Transformer (T5), Quantized Low-Rank Adaptation (Q-LoRA)*

# INTISARI

Perkembangan di bidang *Natural Language Processing* (NLP) saat ini mengalami peningkatan yang luar biasa. Salah satunya dapat dilihat pada *task Question Answering* (QA), yang memungkinkan pengguna berinteraksi dengan sistem secara lebih mudah. Dalam pengembangan sistem QA, salah satu tantangan yang dihadapi adalah bagaimana sistem dapat memberikan jawaban yang akurat sekaligus memperhatikan konteks. Pada sistem QA berbahasa Indonesia yang termasuk dalam kategori *low resource language*, tantangan lainnya adalah keterbatasan pada *dataset* yang tersedia. Hal ini mempengaruhi kinerja model terutama ketika menghadapi variasi dan kompleksitas bahasa yang tidak selalu terwakili dalam *dataset.*

Salah satu solusi untuk mengatasi keterbatasan ini adalah dengan menerjemahkan dataset dari bahasa *resource-rich* (seperti bahasa Inggris) ke bahasa Indonesia. Penggunaan model Transformer, seperti T5, dalam pengembangan sistem QA telah menjadi fokus penelitian terkait. Model T5 memungkinkan penerapan *transfer learning* tanpa perlu penyesuaian arsitektur model secara signifikan dan format *text-to-text* memberikan fleksibilitas yang besar dalam menyesuaikan model dengan berbagai domain dan jenis tugas NLP termasuk QA. Sayangnya, kendala utama dari model T5 adalah kebutuhan akan sumber daya komputasi yang besar, karena jumlah parameternya mencapai 246.078.720. Hal ini menjadi hambatan, terutama bagi pengguna yang tidak memiliki akses ke komputasi tingkat tinggi atau infrastruktur memadai. Oleh karena itu, penelitian ini bertujuan untuk mengkompresi jumlah parameter model T5 agar dapat digunakan pada sumber daya dengan komputasi yang terbatas.

Berdasarkan hasil eksperimen, penerapan Q-LoRA pada proses *fine-tuning* dengan model T5 berhasil mengurangi jumlah parameter menjadi 1.769.472, jauh lebih rendah dibandingkan dengan model awal T5. Selain itu, penerapan Q-LoRA juga menghasilkan peningkatan efisiensi waktu pengujian. Langkah ini tidak hanya mengurangi kebutuhan memori tetapi juga meningkatkan efisiensi komputasi sehingga dapat digunakan pada sumber daya dengan komputasi yang terbatas.

**Kata Kunci:** *Natural Language Processing*, *Question Answering*, Metode Kompresi, *Text-to-Text Transfer Transformer* (T5), *Quantized Low-Rank Adaptation* (Q-LoRA)