

INTISARI

Penelitian ini bertujuan membandingkan sembilan Small Language Model (SLM) lintas keluarga arsitektur berukuran 270 juta hingga dua miliar parameter untuk tugas Natural Language to Function Calling (NL2FC) berbahasa Indonesia pada domain kepegawaian Aparatur Sipil Negara (ASN). Penelitian dilaksanakan melalui beberapa tahapan berurutan. Dataset sebanyak 1.150 sampel yang mencakup sembilan fungsi kepegawaian dibangkitkan dan divalidasi menggunakan LLM API, kemudian dibagi secara berstrata dengan rasio 70:15:15. Kesembilan model selanjutnya diadaptasi secara independen menggunakan Low-Rank Adaptation (LoRA) dengan konfigurasi hyperparameter yang seragam, lalu dievaluasi pada kondisi zero-shot dan fine-tuned. Pelatihan dilaksanakan pada GPU NVIDIA A100, sedangkan pengujian dijalankan pada GPU NVIDIA Tesla T4. Kinerja tiap model diukur menggunakan Function Name Accuracy (FNA), Parameter Accuracy, Exact Match Accuracy, Precision, Recall, dan F1-Score makro per fungsi, serta latensi inferensi, sebelum hasil kesembilan model dibandingkan untuk merumuskan rekomendasi arsitektur.

Hasil penelitian menunjukkan bahwa fine-tuning menjadi faktor penentu kinerja. Lima model meraih F1-Score makro di atas 93%, dengan Qwen3.5-0.8B dan Qwen3.5-2B tertinggi pada FNA 98,3%. FunctionGemma-270M memperoleh peningkatan terbesar, hampir 59 poin F1, sedangkan Llama 3.2-1B, DeepSeek-R1-Distill-1.5B, dan Qwen2.5-Coder-1.5B gagal atau tidak stabil. Latensi terbukti bersifat arsitektural; Qwen3.5-0.8B memadukan F1-Score tertinggi (98,0%) dengan latensi 3.809 milidetik per sampel pada T4, hampir tiga kali lebih cepat daripada Qwen3.5-2B yang akurasinya setara.

Disimpulkan bahwa SLM yang diadaptasi pada domain layak menjadi komponen tool-routing sistem informasi kepegawaian berbasis bahasa natural, dengan Qwen3.5-0.8B sebagai model yang direkomendasikan dan telah terbukti layak dioperasikan pada perangkat keras kelas T4 yang terjangkau, serta FunctionGemma-270M dan Granite 4.0-350M sebagai alternatif yang lebih ringkas. Arsitektur yang disarankan memisahkan penafsiran bahasa oleh SLM dari eksekusi operasi data sehingga aman, mudah diaudit, dan dapat dioperasikan secara lokal.

Kata kunci: small language model, function calling, fine-tuning, kepegawaian ASN, Bahasa Indonesia

ABSTRACT

This study aims to compare nine Small Language Models (SLMs) across architecture families, ranging from 270 million to two billion parameters, for Indonesian Natural Language to Function Calling (NL2FC) in the civil servant (ASN) personnel domain. The research was carried out through a sequence of stages. A dataset of 1,150 samples covering nine personnel functions was generated and validated using an LLM API, then split into a stratified 70:15:15 ratio. The nine models were subsequently adapted independently using Low-Rank Adaptation (LoRA) with a uniform hyperparameter configuration, and evaluated under zero-shot and fine-tuned conditions. Training was conducted on an NVIDIA A100 GPU, while testing was carried out on an NVIDIA Tesla T4 GPU. Each model's performance was measured using Function Name Accuracy (FNA), Parameter Accuracy, Exact Match Accuracy, Precision, Recall, and macro F1-Score per function, along with inference latency, before the results of all nine models were compared to formulate an architecture recommendation.

The results show that fine-tuning is the decisive factor in performance. Five models achieved macro F1-Scores above 93%, with Qwen3.5-0.8B and Qwen3.5-2B highest at 98.3% FNA. FunctionGemma-270M gained the largest improvement of nearly 59 F1 points, whereas Llama 3.2-1B, DeepSeek-R1-Distill-1.5B, and Qwen2.5-Coder-1.5B failed or proved unstable. Latency proved to be architectural in nature; Qwen3.5-0.8B combined the highest F1-Score (98.0%) with a latency of 3,809 milliseconds per sample on the T4, nearly three times faster than the equally accurate Qwen3.5-2B.

It is concluded that domain-adapted SLMs are viable as tool-routing components in natural-language-based personnel information systems, with Qwen3.5-0.8B as the recommended model, shown to remain viable on affordable T4-class hardware, and FunctionGemma-270M and Granite 4.0-350M as leaner alternatives. The recommended architecture separates language interpretation by the SLM from data operation execution, keeping the system secure, auditable, and locally deployable.

Keywords: small language model, function calling, fine-tuning, civil servant personnel management, Indonesian language