TESIS

KOMPARASI PERFORMA *LARGE LANGUAGE MODELS* UNTUK TUGAS PERINGKASAN TEKS BERBAHASA INDONESIA

Oleh:

R. Abdullah Hammami

NIM: 22206051019

PROGRAM STUDI INFORMATIKA

PROGRAM MAGISTER FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA

YOGYAKARTA

2025

PERNYATAAN KEASLIAN

Yang bertanda tangan di bawah ini:

Nama	: R. Abdullah Hammami
NIM	: 22206051019
Jenjang	: Magister
Program Studi	: Informatika

menyatakan bahwa naskah tesis ini secara keseluruhan adalah hasil penelitian/ karya saya sendiri, kecuali pada bagian-bagian yang dirujuk sumbernya.

Yogyakarta, 16 Agustus 2025

Saya yang menyatakan,

R. Abdullah Hammami
NIM: 22206051019

PERNYATAAN BEBAS PLAGIASI

Yang bertanda tangan di bawah ini:

Nama	: R. Abdullah Hammami
NIM	: 22206051019
Jenjang	: Magister
Program Studi	: Informatika

menyatakan bahwa naskah tesis ini secara keseluruhan benar-benar bebas dari plagiasi. Jika di kemudian hari terbukti melakukan plagiasi, maka saya siap ditindak sesuai ketentuan hukum yang berlaku.

Yogyakarta, 16 Agustus 2025

Saya yang menyatakan,

R. Abdullah Hammami
NIM: 22206051019

KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA
FAKULTAS SAINS DAN TEKNOLOGI
Jl. Marsda Adisucipto Telp. (0274) 540971 Fax. (0274) 519739 Yogyakarta 55281

PENGESAHAN TUGAS AKHIR

Nomor : B-1992/Un.02/DST/PP.00.9/08/2025

Tugas Akhir dengan judul : Komparasi Performa Large Language Models untuk Tugas Peringkasan Teks Berbahasa Indonesia

yang dipersiapkan dan disusun oleh:

Nama	: R. ABDULLAH HAMMAMI, S.Kom., CADS
Nomor Induk Mahasiswa	: 22206051019
Telah diujikan pada	: Jumat, 22 Agustus 2025
Nilai ujian Tugas Akhir	: A/B

dinyatakan telah diterima oleh Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta

TIM UJIAN TUGAS AKHIR

Ketua Sidang

Dr. Agung Fatwanto, S.Si., M.Kom.
SIGNED

Valid ID: 68ad8b34ec612

Penguji I

Muhammad Mustakim, S.T. M.T.
SIGNED

Valid ID: 68ad7c1564df8

Penguji II

Dr. Ir. Sumarsono, S.T., M.Kom.
SIGNED

Valid ID: 68ad790380b5f

Yogyakarta, 22 Agustus 2025
UIN Sunan Kalijaga
Dekan Fakultas Sains dan Teknologi

Prof. Dr. Dra. Hj. Khurul Wardati, M.Si.
SIGNED

Valid ID: 68ac87a04994f

NOTA DINAS PEMBIMBING

Kepada Yth.,
Dekan Fakultas Sains dan Teknologi
UIN Sunan Kalijaga
Yogyakarta

Assalamu'alaikum wr. wb.

Setelah melakukan bimbingan, arahan, dan koreksi terhadap penulisan tesis yang berjudul:

KOMPARASI PERFORMA *LARGE LANGUAGE MODELS* UNTUK TUGAS PERINGKASAN TEKS BERBAHASA INDONESIA

Yang ditulis oleh:

Nama	: R. Abdullah Hammami
NIM	: 22206051019
Jenjang	: Magister
Program Studi	: Informatika

Saya berpendapat bahwa tesis tersebut sudah dapat diajukan kepada Magister Informatika UIN Sunan Kalijaga untuk diujikan dalam rangka memperoleh gelar Magister Komputer.

Wassalamu'alaikum wr. wb.

Yogyakarta, 16 Agustus 2025

Pembimbing

Dr. Agung Fatwanto, S.Si., M.Kom.
NIP. 19770103 200501 1 003

ABSTRAK

Arus informasi daring yang masif, rendahnya minat baca, dan beragamnya tingkat literasi di Indonesia menuntut peringkasan otomatis yang ringkas, akurat, dan peka konteks. Karena Bahasa Indonesia tergolong low-resource, perlu evaluasi sistematis terhadap model yang telah disesuaikan secara lokal. Studi ini membandingkan empat LLM berbahasa Indonesia, yaitu Gemma2 9B CPT Sahabat-AI v1 Instruct, Llama3 8B CPT Sahabat-AI v1 Instruct, Gemma-SEA-LION-v3-9B-IT, dan Llama-SEA-LION-v3-8B-IT pada tugas peringkasan berita. Metode berupa benchmarking di subset uji IndoSum (3.762 artikel), mencakup pra-pemrosesan (rekonstruksi token, sanitasi tanda baca), perancangan prompt, inferensi berkuantisasi 8-bit, serta evaluasi ROUGE (1/2/L; precision, recall, F1), BLEU, METEOR, dan BERTScore. Hasil: Llama3 8B CPT Sahabat-AI v1 Instruct paling seimbang dengan ROUGE F1 42,05% (precision 42,27%; recall 42,68%), BLEU 25,10%, BERTScore P/R/F1 88,68%/88,43%/88,54%. Gemma2 9B CPT Sahabat-AI v1 Instruct unggul pada cakupan: ROUGE recall 48,23%, ROUGE F1 39,50%, BLEU 22,70%, METEOR 47,20%, BERTScore 86,78%/89,17%/87,95%. Model SEA-LION lebih rendah: Gemma-SEA-LION-v3-9B-IT (ROUGE P/R/F1 25,77%/37,58%/30,37%; BLEU 12,65%; METEOR 37,72%; BERTScore 84,63%/87,36%/85,97%) dan Llama-SEA-LION-v3-8B-IT (ROUGE 25,22%/33,84%/28,71%; BLEU 11,06%; METEOR 34,57%; BERTScore 84,46%/86,80%/85,61%). Model SahabatAI tampil lebih unggul dan stabil; Llama3 8B cocok untuk keseimbangan presisi, cakupan, dan konsistensi struktur, sedangkan Gemma2 9B lebih tepat saat prioritas pada recall dan kesepadanan makna dengan sumber.

Kata kunci: Peringkasan Teks, Fine-tuning, Gemma2, LLaMA3

ABSTRACT

The rapid growth of online information, coupled with low reading interest and heterogeneous literacy levels in Indonesia, necessitates concise, accurate, and context-sensitive automatic summarization. Given Indonesian's low-resource status, systematic evaluation of locally adapted models is warranted. This study compares four Indonesian-capable large language models—Gemma2 9B CPT Sahabat-AI v1 Instruct, Llama3 8B CPT Sahabat-AI v1 Instruct, Gemma-SEA-LION-v3-9B-IT, and Llama-SEA-LION-v3-8B-IT—on news summarization to identify the most suitable model for practical use. We employ a benchmarking protocol on the IndoSum test subset (3,762 articles), comprising preprocessing (token reconstruction and punctuation cleanup), prompt design, 8-bit quantized inference, and automated evaluation with ROUGE (1/2/L; precision, recall, F1), BLEU, METEOR, and BERTScore. Inference is executed in four batches to meet computational constraints, and evaluation is standardized across models. Llama3 8B CPT Sahabat-AI v1 Instruct achieves the most balanced performance: ROUGE F1 42.05% (precision 42.27%; recall 42.68%), BLEU 25.10%, and BERTScore P/R/F1 88.68%/88.43%/88.54%. Gemma2 9B CPT Sahabat-AI v1 Instruct excels in coverage with ROUGE recall 48.23%, ROUGE F1 39.50%, BLEU 22.70%, METEOR 47.20%, and BERTScore 86.78%/89.17%/87.95%. SEA-LION models perform lower: Gemma-SEA-LION-v3-9B-IT (ROUGE P/R/F1 25.77%/37.58%/30.37%; BLEU 12.65%; METEOR 37.72%; BERTScore 84.63%/87.36%/85.97%) and Llama-SEA-LION-v3-8B-IT (ROUGE 25.22%/33.84%/28.71%; BLEU 11.06%; METEOR 34.57%; BERTScore 84.46%/86.80%/85.61%). Overall, Indonesian-optimized models (SahabatAI) are superior and more stable. Llama3 8B is preferable when balancing precision, coverage, and structural consistency; Gemma2 9B is better when recall and semantic alignment with the source are prioritized.

Keywords: Text Summarization, Fine-tuning, Gemma2, LLaMA3

KATA PENGANTAR

Bissmillahirahmanirrahim

Assalamu'alaikum wa rahmatullahi wa barakatuh

Alhamdulillah alladzi qod wafaqo lil 'ilmi khoiro kholqihi wa littuqo. Puji syukur ke hadirat Allah subhanahu wa ta'ala atas segala rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan tesis yang berjudul "*Komparasi Performa Large Language Models untuk Tugas Peringkasan Teks Berbahasa Indonesia*." Tesis ini disusun sebagai salah satu syarat dalam menyelesaikan program pendidikan pada jenjang magister, sekaligus sebagai bentuk kontribusi ilmiah dalam bidang pemrosesan bahasa alami (Natural Language Processing/NLP) dengan fokus pada bahasa Indonesia. Penulis menyadari bahwa perkembangan teknologi kecerdasan buatan, khususnya *Large Language Models* (LLM), menghadirkan peluang besar bagi peningkatan kualitas sistem peringkasan teks otomatis. Oleh karena itu, penelitian ini diharapkan dapat memberikan manfaat bagi pengembangan teknologi AI yang relevan dengan konteks linguistik dan budaya Indonesia. Tak lupa penulis mengucapkan terima kasih kepada semua pihak yang terlibat dan berkontribusi dalam penyusunan tesis ini. Untaian terima kasih tersebut penulis tujukan kepada:

1. Kementerian Agama Republik Indonesia selaku pemberi beasiswa PBSB.
2. Bapak Prof. Noorhaidi, M.A., M.Phil., Ph.D., selaku Rektor Universitas Islam Negeri Sunan Kalijaga Yogyakarta
3. Ibu Dr. Dra. Hj. Khurul Wardati, M.Si., selaku Dekan Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta
4. Bapak Dr. Ir. Sumarsono, S.T., M.Kom., selaku Ketua Program Studi Magister Informatika Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta
5. Bapak Dr. Agung Fatwanto, S.Si., M.Kom., selaku dosen Pembimbing Tesis yang telah memberikan arahan selama studi sehingga dapat menyelesaikan penyusunan tesis ini

6. Bapak Ibu dosen dan staf akademik Program Studi Magister Informatika Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta yang telah membimbing dan memberikan ilmu yang semoga dapat bermanfaat
7. Orang tua penulis terutama Ibu yang selalu dan senantiasa memberikan support dan doa kepada penulis selama menempuh pendidikan Magister Informatika UIN Sunan Kalijaga Yogyakarta.
8. Dan terakhir kepada seluruh teman-teman yang mungkin tidak bisa penulis sebutkan satu persatu, baik dari teman-teman angkatan 2022, PBSB dan lainnya yang telah men-support penulis baik semoga Allah membalas kebaikan teman-teman semua.

Akhir kata, penulis menyadari sepenuhnya bahwa tesis ini masih jauh dari sempurna. Oleh karena itu, kritik dan saran yang membangun sangat penulis harapkan demi penyempurnaan penelitian ini di masa mendatang. Penulis juga berharap semoga karya ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan, khususnya dalam bidang pemrosesan bahasa alami dan penerapan teknologi kecerdasan buatan di Indonesia. Semoga tesis ini dapat menjadi referensi yang berguna bagi penelitian selanjutnya serta bermanfaat bagi semua pihak yang berkecimpung dalam dunia akademik maupun praktisi.

Yogyakarta, 16 Agustus 2025

R. Abdullah Hammami

PERSEMBAHAN

Persembahan: Karya ini penulis persembahkan kepada Ibunda Nyimas Maimunah tercinta atas doa, motivasi dan kasih tanpa batas; kepada almarhum Ayahanda RM. Abdul Hakim, semoga Allah senantiasa merahmatinya dan seluruh keluarga yang menjadi penopang setiap langkah; serta kepada almamater UIN Sunan Kalijaga, tempat bernaung dan bertumbuhnya ilmu dan akhlak, sebagai wujud syukur dan bakti atas setiap perjuangan yang mengantarkan penelitian ini hingga tuntas.

MOTTO

إِذِ الْفَتَى حَسْبَ اعْتِقَادِهِ رُفِعْ # وَكُلُّ مَنْ لَمْ يَعْتَقِدْ لَم يَنْتَفِعْ

(*Ketika keyakinan seseorang kuat maka akan diangkat derajatnya # dan setiap insan yang tidak memiliki keyakinan maka tidak akan bisa mengambil manfaat*)

– Ibnu Malik –

(dalam Alfiyah Ibnu Malik)

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I
PENDAHULUAN

A. LATAR BELAKANG

Perkembangan pesat dalam bidang kecerdasan buatan (Artificial Intelligence/AI) telah membawa perubahan signifikan terhadap cara manusia mengakses, memproses, dan memahami informasi. Teknologi AI kini banyak digunakan dalam berbagai aktivitas sehari-hari, seperti menerjemahkan berita atau artikel secara otomatis, memberikan rekomendasi konten yang sesuai minat, mendeteksi berita palsu (hoaks), mengelompokkan dokumen berdasarkan topik, hingga membuat ringkasan otomatis dari teks yang panjang. Salah satu bidang yang semakin mendapat perhatian adalah peringkasan teks otomatis, yang bertujuan membantu pengguna memperoleh inti sari suatu dokumen dengan cepat dan efisien.

Di era digital, arus informasi yang sangat deras memunculkan fenomena *information overload*, terutama pada konsumsi berita daring, di mana ribuan artikel baru dipublikasikan setiap hari. Keterbatasan waktu dan kapasitas manusia dalam menyaring serta memahami konten mendorong kebutuhan akan sistem peringkasan otomatis yang tidak hanya ringkas, tetapi juga akurat dan mampu mempertahankan informasi penting dari teks sumber. Tantangan ini semakin relevan di Indonesia, mengingat rendahnya minat baca masyarakat. Data UNESCO menunjukkan bahwa minat baca masyarakat Indonesia hanya sekitar 0,001%, yang berarti dari 1.000 orang, hanya 1 orang yang gemar membaca. Hasil survei PISA 2017 juga menempatkan Indonesia pada peringkat 62 dari 72 negara dalam literasi (Kurniati dkk., 2023).

Rendahnya minat baca ini berdampak pada kemampuan memahami informasi dari teks. Berdasarkan data *Indonesian National Assessment Program*, hanya 6,06% siswa Indonesia yang memiliki kemampuan membaca baik, sementara 47,11% berada pada kategori sedang, dan 46,83% pada kategori rendah . Penelitian di SMAN 10 Pekanbaru juga menunjukkan bahwa meskipun minat baca siswa rata-rata berada pada tingkat sedang (*mean* 70) dan kemampuan memahami bacaan berada pada tingkat baik (*mean* 77), terdapat korelasi positif dengan kekuatan sedang ($r = 0,423$, $p = 0,018$) antara keduanya (Kurniati dkk., 2023). Artinya, semakin tinggi minat baca, semakin baik pula kemampuan memahami informasi.

Kondisi rendahnya minat baca dan variasi tingkat pemahaman ini menjadi tantangan sekaligus landasan penting dalam pengembangan sistem peringkasan teks untuk bahasa Indonesia. Sistem peringkasan tidak hanya dituntut mampu menyaring informasi secara akurat, tetapi juga menyajikannya dalam bentuk ringkasan yang jelas, sederhana, dan mudah dipahami oleh pembaca dengan beragam tingkat literasi. Tantangan ini semakin kompleks mengingat bahasa Indonesia memiliki keragaman struktur kalimat, variasi gaya penulisan, penggunaan istilah lokal, serta konteks budaya yang khas, yang membedakannya dari bahasa Inggris atau bahasa mayoritas dunia lainnya. Selain itu, bahasa Indonesia tergolong *low-resource language* karena ketersediaan data pelatihan dan sumber daya pendukung yang relatif terbatas. Oleh sebab itu, *Large Language Models (LLM)* perlu mendapatkan penyesuaian dan pelatihan khusus agar dapat menghasilkan ringkasan yang relevan, akurat, serta mudah dipahami sesuai karakteristik sintaksis dan semantik bahasa Indonesia.

Berbagai penelitian sebelumnya telah mencoba pendekatan ekstraktif seperti BERT (Halim dkk., 2022) dan Algoritma *Ant System* (Girsang & Amadeus, 2023), serta metode abstraktif berbasis transformer seperti T5 (Itsnaini dkk., 2023), BART (Harditya & Purnamasari, 2024), GPT-2 (Khasanah & Hayaty, 2023), dan metode *Sequance-to-sequance (Seq2seq)* dengan LSTM dan PEGASUS (Batari, 2023). Namun, kesenjangan performa antara hasil peringkasan berbahasa Indonesia dan bahasa mayoritas lain masih terlihat signifikan, baik dari segi akurasi, koherensi, maupun kelengkapan informasi. Sebagai contoh, sebuah penelitian yang mengevaluasi enam model LLM open-source (LLaMA-2, Mistral-7B-Instruct, Gemma-7B-it, Mixtral-8x7B-Instruct) pada empat dataset (berita, dialog, ilmiah), yang menunjukkan bahwa meskipun semua model kompetitif, variasi performa sangat dipengaruhi oleh jenis model, dataset, dan desain prompt (Rehman dkk., 2025).

Penelitian lain yang dipublikasikan di Arxiv membandingkan performa beberapa model LLM seperti BART, FLAN-T5, LLaMA-3-8B, dan Gemma-7B di berbagai domain dan dataset, menggunakan metrik ROUGE, METEOR, dan BERT*Score*. Hasilnya menunjukkan bahwa Gemma-7B sangat informatif, konsisten, dan kompetitif, meskipun terdapat variasi performa antar dataset (Aly dkk., 2025). Selain itu, pengembangan model Gemma 2, yang memiliki berbagai ukuran parameter (2B–27B), memperkenalkan peningkatan efisiensi dalam arsitektur, serta teknik *fine-tuning* dan *knowledge distillation*, yang memungkinkan model ini unggul pada berbagai tugas NLP termasuk peringkasan teks (Team, Riviere, dkk., 2024).

Berdasarkan permasalahan tersebut, penelitian ini difokuskan pada evaluasi dan perbandingan kinerja beberapa model LLM yang telah melalui proses *pretraining* dan *fine-tuning* khusus untuk bahasa Indonesia, yaitu Gemma2 9B CPT Sahabat-AI v1 Instruct, Llama3 8B CPT Sahabat-AI v1 Instruct, Gemma-SEA-LION-v3-9B-IT, dan Llama-SEA-LION-v3-8B-IT. Evaluasi dilakukan menggunakan dataset IndoSum, yang berisi berita harian berbahasa Indonesia beserta ringkasan buatan manusia, dengan penerapan metrik evaluasi standar di bidang peringkasan teks: ROUGE, BLEU, METEOR, dan BERT*Score*. Setiap metrik dipilih untuk mengukur dimensi yang berbeda, mulai dari kesamaan n-gram, kelengkapan informasi, hingga kesesuaian semantik antara ringkasan yang dihasilkan dan referensi.

Hasil penelitian diharapkan dapat memberikan kontribusi signifikan dalam pemilihan model LLM yang paling optimal untuk kebutuhan peringkasan teks bahasa Indonesia dan menjadi referensi dalam pengembangan teknologi peringkasan yang adaptif, efisien, serta relevan dengan konteks dan kebutuhan masyarakat yang ada di Indonesia.

B. RUMUSAN MASALAH

Penelitian ini bertujuan untuk mengevaluasi performa beberapa *Large Language Model* (LLM) yang sering digunakan dalam berbagai tugas, khususnya pada tugas peringkasan teks berbahasa Indonesia. Hingga saat ini, masih belum diketahui model mana yang memiliki performa terbaik. Dalam penelitian ini, model yang dibandingkan adalah **Gemma2 9B CPT Sahabat-AI v1 Instruct, Llama3 8B CPT Sahabat-AI v1 Instruct, Gemma-SEA-LION-v3-9B-IT, dan Llama-**

SEA-LION-v3-8B-IT. Keempat model tersebut telah mendukung bahasa Indonesia karena masing-masing telah melalui proses *pre-training* dan *fine-tuning* menggunakan dataset berbahasa Indonesia dalam jumlah yang cukup besar dari pihak pengembang.

C. BATASAN MASALAH

Agar penelitian ini memiliki arah yang sesuai dengan tujuan utama penelitian ini. Berikut adalah beberapa Batasan masalah yang dimiliki penelitian ini.

1. Penelitian ini dibatasi pada evaluasi model bahasa besar **Gemma2 9B CPT Sahabat-AI v1 Instruct, Llama3 8B CPT Sahabat-AI v1 Instruct, Gemma-SEA-LION-v3-9B-IT** dan **Llama-SEA-LION-v3-8B-IT** saja, tanpa membandingkannya dengan model bahasa besar lain yang ada di luar model-model tersebut.
2. Dataset yang digunakan adalah dataset Indosum yang diunduh melalui laman repositori Github kata-ai.
3. Bahasa yang menjadi rujukan dalam penelitian ini adalah Bahasa Indonesia.
4. Evaluasi performa yang dilakukan dalam penelitian ini menggunakan metrik evaluasi ROUGE (ROUGE-1, ROUGE-2 dan ROUGE-L), BLEU, METEOR, dan BERT*Score* (*Recall, Precission* dan *F1-Score*).

D. TUJUAN PENELITIAN

Berdasarkan uraian rumusan masalah dan batasan masalah yang diatas, peneltian ini ditujukan untuk mengetahui performa dari *Large Language Model* (LLM) yang terbaik dalam tugas peringkasan teks,

yaitu **Gemma2 9B CPT Sahabat-AI v1 Instruct, Llama3 8B CPT Sahabat-AI v1 Instruct, Gemma-SEA-LION-v3-9B-IT** dan **Llama-SEA-LION-v3-8B-IT** dengan menggunakan metrik evaluasi ROUGE, BLEU, METEOR dan BERT*Score* (*Recall, Precission* dan *F1-Score*).

E. MANFAAT PENELITIAN

Dengan melakukan penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Penelitian ini memberikan wawasan mendalam tentang bagaimana LLM *(Large Language Models)* yang telah melalui proses *pre-train* dan *fine-tune* menggunakan bahasa Indonesia, khususnya untuk melakukan tugas peringkasan teks dalam bahasa Indonesia.
2. Dengan menggunakan dataset yang sudah dikurasi oleh manusia, penelitian ini membantu mengidentifikasi sejauh mana performa model bahasa besar **Gemma2 9B CPT Sahabat-AI v1 Instruct, Llama3 8B CPT Sahabat-AI v1 Instruct, Gemma-SEA-LION-v3-9B-IT dan Llama-SEA-LION-v3-8B-IT**.
3. Penelitian ini mengevaluasi dengan berbagai metrik evaluasi seperti ROUGE (ROUGE-1, ROUGE-2 dan ROUGE-L), BLEU, METEOR dan BERT*Score* (*Recall, Precission* dan *F1-Score*) untuk menilai kualitas dari ringkasan teks yang dihasilkan.
4. Dengan melakukan penelitian ini dapat membuka jalan bagi pengembangan untuk implementasi peringkasan teks yang lebih akurat dan relevan dalam bahasa Indonesia, yang masih jarang dijelajahi dibandingkan dengan bahasa-bahasa lain.

5. Dalam penelitian ini menyediakan panduan praktis tentang bagaimana model bahasa besar yang telah melaui proses *pre-train* dan *fine-tune* dapat berpengaruh pada hasil ringkasan teks, yang dapat sangat berguna bagi pengembang model AI (*Artificial Intellegent*) dan praktisi dalam menyesuaikan model untuk kebutuhan yang lebih spesifik khususnya meringkas teks berita.
6. Memberikan gambaran kepada masyarakat khususnya para pengembang aplikasi agar bisa memilih model mana yang memiliki presisi dan akurasi yang bagus dalam tugas peringkasan teks.

F. KEASLIAN PENELITIAN

Dalam melakukan peninjauan literatur dan mengobservasi beberapa penelitian, penelitian yang berkaitan dengan evaluasi performa LLM *(Large Language Models)* **Gemma2 9B CPT Sahabat-AI v1 Instruct, Llama3 8B CPT Sahabat-AI v1 Instruct, Gemma-SEA-LION-v3-9B-IT** dan **Llama-SEA-LION-v3-8B-IT** menggunakan metrik evaluasi ROUGE (ROUGE-1, ROUGE-2 dan ROUGE-L), BLEU, METEOR dan BERT*Score* (*Recall, Precission* dan *F1-Score*) dengan dataset IndoSum yang berbahasa Indonesia yang mana telah dikurasi masih belum pernah dilakukan sebelumnya.

BAB V
PENUTUP

A. KESIMPULAN

Berdasarkan keseluruhan hasil evaluasi, dapat disimpulkan bahwa model dari keluarga SahabatAI, khususnya Llama3 8B CPT Sahabat-AI v1 Instruct, menunjukkan performa paling seimbang di berbagai metrik evaluasi. Model ini mencatat skor tertinggi pada *F1-Score* ROUGE (42,05%), BLEU (25,10%), serta performa BERT*Score* yang sangat stabil (*Precision* 88,68%, *Recall* 88,43%, *F1* 88,54%), menandakan kemampuannya menghasilkan ringkasan yang relevan, akurat, dan menyerupai referensi dari segi struktur maupun pilihan kata. Sementara itu, Gemma2 9B CPT Sahabat-AI v1 Instruct unggul pada *Recall* ROUGE (48,23%) dan METEOR (47,20%), serta mencatat *Recall* BERT*Score* tertinggi (89,17%), yang menunjukkan keunggulan dalam cakupan informasi meskipun presisi masih di bawah model Llama3 SahabatAI.

Di sisi lain, model dari keluarga SEA-LION memperlihatkan performa yang konsisten berada di bawah model SahabatAI pada hampir semua metrik. Gemma-SEA-LION-v3-9B-IT memiliki *Recall* ROUGE yang relatif lebih baik (37,58%) dibandingkan Llama-SEA-LION, namun skor BLEU (12,65%) dan METEOR (37,72%) menunjukkan adanya perbedaan signifikan dengan referensi, terutama dari segi struktur kalimat dan pemilihan kata. Model Llama-SEA-LION-v3-8B-IT menjadi yang terendah di seluruh metrik, dengan *F1* ROUGE hanya 28,71%, BLEU 11,06%, dan METEOR 34,57%, serta skor BERT*Score*

yang meskipun stabil (*Precision* 84,46%, *Recall* 86,80%, *F1* 85,61%), masih tertinggal dari model lain.

Secara umum, hasil ini menegaskan bahwa model yang dioptimalkan khusus untuk bahasa Indonesia, seperti SahabatAI, mampu memberikan kinerja lebih unggul dalam hal kesesuaian semantik, kelengkapan informasi, serta kemiripan n-gram terhadap referensi. Hal ini juga menunjukkan pentingnya keseimbangan antara *precision* dan *recall* untuk menghasilkan ringkasan yang tidak hanya akurat, tetapi juga komprehensif. Lebih jauh lagi, dengan kondisi rendahnya literasi digital di Indonesia serta tantangan *information overload* di era digital, sistem peringkasan otomatis yang akurat, sederhana, dan mudah dipahami menjadi solusi strategis dalam meningkatkan akses informasi yang efisien bagi masyarakat. Oleh karena itu, hasil evaluasi ini dapat dijadikan pertimbangan penting bagi para pengembang aplikasi *text summarization* (peringkasan teks) dalam memilih LLM yang paling presisi dan akurat untuk kebutuhan praktis.

B. SARAN

Berdasarkan hasil evaluasi yang menunjukkan keunggulan model SahabatAI, khususnya Llama3 8B CPT Sahabat-AI v1 Instruct, dalam menghasilkan ringkasan teks berbahasa Indonesia yang relevan dan komprehensif, berikut beberapa saran penelitian selanjutnya yang layak dipertimbangkan.

1. Pengembangan versi Gemma dan LLaMA terbaru atau LLM lainnya dengan melakukan *pre-train* dan *fine-tune*.
2. Menguji LLM yang telah diuji dalam penelitian ini dengan dataset yang memiliki domain berbeda.

3. Menguji keempat LLM untuk melakukan peringkasan terhadap bahasa daerah, seperti bahasa Jawa, Sunda dan lainnya.
4. Memaksimalkan seluruh dataset untuk dievaluasi, sekitar 18 ribu baris (bukan hanya 3672 baris data test).

DAFTAR PUSTAKA

Aly, W. M., Soliman, T. H. A., & AbdelAziz, A. M. (2025). *An Evaluation of Large Language Models on Text Summarization Tasks Using Prompt Engineering Techniques*. https://arxiv.org/abs/2507.05123v1

Ardyanti, C. P. R., Wibisono, Y., & Megasari, R. (2024). Peringkatas Teks Berita Berbahasa Indonesia Menggunakan LSTM dan Transformer. *Indonesian Journal of Applied Informatics*, *2*. https://jurnal.uns.ac.id/ijai/article/view/78856/pdf

Banerjee, S., & Lavie, A. (t.t.). *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*.

Banner, R., Nahshan, Y., & Soudry, D. (2018). Post-training 4-bit quantization of convolution networks for rapid-deployment. *Advances in Neural Information Processing Systems*, *32*. https://arxiv.org/pdf/1810.05723

Batari, R. (2023). *ABSTRACTIVE TEXT SUMMARIZATION PADA ULASAN BUKU BERBAHASA INDONESIA MENGGUNAKAN METODE SEQUENCE-TO-SEQUENCE LSTM DAN PEGASUS* [Skripsi, Universitas Padjadjaran]. https://repository.unpad.ac.id/handle/kandaga/140810180051

Bhandare, A., Sripathi, V., Karkada, D., Menon, V., Choi, S., Datta, K., & Saletore, V. (2019). *Efficient 8-Bit Quantization of Transformer Neural Machine Language Translation Model*. https://arxiv.org/pdf/1906.00532

Bird, S., & Loper, E. (t.t.). *NLTK: The Natural Language Toolkit*. Diambil 30 Juli 2025, dari www.python.org.

Brazdil, P., van Rijn, J. N., Soares, C., & Vanschoren, J. (2022). Dataset Characteristics (Metafeatures). *Cognitive Technologies*, 53–75. https://doi.org/10.1007/978-3-030-67024-5_4

Carosso, A. (2022). Quantization: History and problems. *Studies in History and Philosophy of Science*, *96*, 35–50. https://doi.org/10.1016/J.SHPSA.2022.09.001

Doing Data Science with coLaboratory. (t.t.). Diambil 29 Juli 2025, dari https://research.google/blog/doing-data-science-with-colaboratory/

Fauzi, D., & Abidin, Z. (2024). Peringkasan Teks Multi-Dokumen Berbahasa Indonesia Dengan Sentence Scoring dan SVM Multi-Dokumen Text Summarization in Indonesian Using Sentence Scoring and SVM. *Techno.COM*, *23*(1), 187–197.

Girsang, A. S., & Amadeus, F. J. (2023). Extractive Text Summarization for Indonesian News Article Using Ant System Algorithm. *Journal of Advances in Information Technology*, *14*(2), 295–301. https://doi.org/10.12720/jait.14.2.295-301

Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L. (2023). A survey on dataset quality in machine learning. *Information and Software Technology*, *162*, 107268. https://doi.org/10.1016/J.INFSOF.2023.107268

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … Ma, Z. (2024). *The Llama 3 Herd of Models*. https://arxiv.org/pdf/2407.21783

Halim, F., Liliana, & Gunadi, K. (2022). Ringkasan Ekstraktif Otomatis pada Berita Berbahasa Indonesia Menggunakan Metode BERT. *Jurnal Infra*.

Harditya, M., & Purnamasari, P. D. (2024). *Pengembangan Abstractive-Extractive Text Summarization dengan BART untuk Teks Berita Bahasa Indonesia* [Skripsi]. Universitas Indonesia.

Hartawan, G., Sa'adillah Maylawati, D., & Uriawan, W. (2024). BIDIRECTIONAL AND AUTO-REGRESSIVE TRANSFORMER (BART) FOR INDONESIAN ABSTRACTIVE TEXT SUMMARIZATION. *JIP (Jurnal Informatika Polinema)*.

Hayashi, T., Shimizu, T., & Fukami, Y. (t.t.). *Collaborative Problem Solving on a Data Platform Kaggle **. Diambil 29 Juli 2025, dari https://www.kaggle.com/kaggle/meta-kaggle

History — Matplotlib 3.10.3 documentation. (t.t.). Diambil 30 Juli 2025, dari https://matplotlib.org/stable/project/history.html

Itsnaini, Q. A., Hayaty, M., Putra, A. D., & Jabari, N. A. M. (2023). Abstractive Text Summarization using Pre-Trained Language Model "Text-to-Text Transfer Transformer (T5)." *ILKOM Jurnal Ilmiah*, *15*(1), 124–131. https://doi.org/10.33096/ilkom.v15i1.1532.124-131

Jin, R., Du, J., Huang, W., Liu, W., Luan, J., Wang, B., & Xiong, D. (2024). *A Comprehensive Evaluation of Quantization Strategies for Large Language Models*. https://arxiv.org/pdf/2402.16775v1

Jones, J., Jiang, W., Synovic, N., Thiruvathukal, G. K., & Davis, J. C. (t.t.). What do we know about Hugging Face? A systematic literature review and quantitative validation of qualitative claims. *ESEM '*, *24*. https://doi.org/10.1145/3674805.3686665

Khasanah, A. N., & Hayaty, M. (2023). ABSTRACTIVE-BASED AUTOMATIC TEXT SUMMARIZATION ON INDONESIAN NEWS USING GPT-2. *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, *10*(1), 9–18. https://doi.org/10.33330/jurteksi.v10i1.2492

Kurniati, R., Daud, A., & Masyhur, M. (2023). Reading Interest and Reading Comprehension Ability: The Correlational Study in Secondary Education. *JELITA: Journal of Education, Language Innovation, and Applied Linguistics*, *2*(1), 41–50. https://doi.org/10.37058/jelita.v2i1.6547

Kurniawan, K., & Louvan, S. (2018). IndoSum: A New Benchmark Dataset for Indonesian Text Summarization. *Proceedings of the 2018 International Conference on Asian Language Processing, IALP 2018*, 215–220. https://doi.org/10.1109/IALP.2018.8629109

Lang, J., Guo, Z., & Huang, S. (2024). A Comprehensive Study on Quantization Techniques for Large Language Models. *2024 4th International Conference on Artificial Intelligence, Robotics, and Communication, ICAIRC 2024*, 224–231. https://doi.org/10.1109/ICAIRC64177.2024.10899941

Lin, C.-Y. (t.t.). *ROUGE: A Package for Automatic Evaluation of Summaries*.

Luhn, H. P. (t.t.). *The Automatic Creation of Literature Abstracts**.

Maylawati, D. S. adillah, Kumar, Y. J., & Kasmin, F. (2023). Feature-based approach and sequential pattern mining to enhance quality of Indonesian automatic text summarization. *Indonesian Journal of Electrical Engineering and Computer Science*, *30*(3), 1795–1804. https://doi.org/10.11591/ijeecs.v30.i3.pp1795-1804

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (t.t.). *Large Language Models: A Survey*.

Naveed, H., Ullah Khan, A., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (t.t.). *A Comprehensive Overview of Large Language Models*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (t.t.). *BLEU: a Method for Automatic Evaluation of Machine Translation*.

Pérez, F., & Granger, B. E. (2007). IPython: A system for interactive scientific computing. *Computing in Science and Engineering*, *9*(3), 21–29. https://doi.org/10.1109/MCSE.2007.53

Poincaré, H. (1912). Sur la théorie des quanta. *Journal de Physique Théorique et Appliquée*, *2*(1), 5–34. https://doi.org/10.1051/jphystap:0191200200500

Pol, U. R., Vadar, P. S., & Moharekar, T. T. (2024). *Hugging Face: Revolutionizing AI and NLP*. *12*. https://doi.org/10.22214/ijraset.2024.64023

Rehman, T., Ghosh, S., Das, K., Bhattacharjee, S., Sanyal, D. K., & Chattopadhyay, S. (2025). *Evaluating LLMs and Pre-trained Models for Text Summarization Across Diverse Datasets*. https://arxiv.org/pdf/2502.19339

Renear, A. H., Sacchi, S., & Wickett, K. M. (2010). Definitions of dataset in the scientific and technical literature. *Proceedings of the ASIST Annual Meeting*, *47*(1), 1–4. https://doi.org/10.1002/MEET.14504701240;PAGE:STRING:ARTICLE/CHAPTER

Sahabat-AI | Open-Source LLMs for Bahasa Indonesia. (t.t.). Diambil 30 Juli 2025, dari https://sahabat-ai.com/#about

Shane Greenstein, Daniel Yue, Kerry Herman, & Sarah Gulick. (2022). *Hugging Face: Serving AI on a platform*. 1–25. https://www.hbs.edu/faculty/Pages/item.aspx?num=63185

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Sessa, P. G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., … Kenealy, K. (2024). *Gemma: Open Models Based on Gemini Research and Technology*. https://arxiv.org/pdf/2403.08295

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. Le, Jerome, S., … Andreev, A. (2024). *Gemma 2: Improving Open Language Models at a Practical Size*. https://arxiv.org/pdf/2408.00118

Thiyagalingam, J., Shankar, M., Fox, G., & Hey, T. (2022). Scientific machine learning benchmarks. *Nature Reviews Physics*, *4*(6), 413–420. https://doi.org/10.1038/s42254-022-00441-7

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models*. https://arxiv.org/pdf/2302.13971

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. https://arxiv.org/pdf/2307.09288

van Rossum, G., & de Boer, J. (1991). Interactively testing remote servers using the Python programming language. *CWI Quarterly*, *4*(4), 283–304.

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2023). *Attention Is All You Need*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. *8th International Conference on Learning Representations, ICLR 2020*. https://arxiv.org/pdf/1904.09675