Evaluasi Teknik Prompt Engineering pada LLM Berukuran Kecil untuk Klasifikasi Sentimen Bahasa Indonesia

Rita Rahmawati [1], Aliya Shidqina Salsabila [2]

[1] *S1 Teknologi Informasi, Fakultas Sains dan Humaniora, Universitas Muhammadiyah Gombong, Kebumen, Indonesia*
[2] *S1 Teknologi Informasi, Fakultas Informatika, Universitas Telkom, Bandung, Indonesia*
Email: *ritarahmawati@unimugo.ac.id

Abstract **—Sentiment classification is a widely used natural language processing task for analyzing public opinion from Indonesian-language text, whether from social media or product reviews. The emergence of Large Language Models (LLMs) has opened new opportunities for performing sentiment classification without training a model from scratch, requiring only the design of an appropriate prompt. However, LLM performance is strongly influenced by the prompting technique used, and research comparing prompting techniques for Indonesian-language text remains limited. This study aims to evaluate and compare the effectiveness of prompt engineering techniques, namely zero-shot, few-shot, and chain-of-thought, in the task of Indonesian-language sentiment classification. Experiments were conducted on 200 test samples from the SmSA (IndoNLU) dataset using two small open-source models, Qwen3-4B-Instruct-2507 and Llama-3.2-3B-Instruct, evaluated on accuracy, precision, recall, and macro F1-score, with Wilson confidence intervals, paired McNemar tests, and a majority-class baseline. All configurations substantially outperformed the majority baseline (accuracy 0.460; F1 0.210). An audit of label extraction revealed that the initial implementation had mis-parsed 18 of 200 outputs in one condition (systematically and in one direction) despite that condition reporting a 0% unknown rate; correcting this raised its accuracy from 0.855 to 0.945. The best configuration was few-shot prompting on Qwen3-4B-Instruct-2507 (accuracy 0.945; macro F1 0.918). Qwen3-4B-Instruct-2507 significantly outperformed Llama-3.2-3B-Instruct under all three techniques (exact McNemar with Holm correction). The only significant difference between prompting techniques was few-shot outperforming chain-of-thought on Qwen3-4B-Instruct-2507 (Holm p=0.038); chain-of-thought was never the best technique on either model while costing tens of times more inference time. All differences between conditions originated from the neutral class, which both models handled far less reliably than positive and negative sentiment.**

Keywords— prompt engineering, large language model, sentiment classification, Indonesian language, zero-shot, few-shot

Abstrak **— Klasifikasi sentimen merupakan salah satu tugas pemrosesan bahasa alami yang banyak digunakan untuk menganalisis opini publik dari teks berbahasa Indonesia, baik dari media sosial maupun ulasan produk. Kemunculan Large Language Model (LLM) membuka peluang baru untuk melakukan klasifikasi sentimen tanpa memerlukan pelatihan model dari awal, cukup melalui perancangan prompt yang tepat. Namun, performa LLM sangat dipengaruhi oleh teknik prompt yang digunakan, dan penelitian yang berfokus pada perbandingan teknik prompt untuk teks berbahasa Indonesia masih terbatas. Penelitian ini bertujuan untuk mengevaluasi dan membandingkan efektivitas teknik prompt engineering, yaitu zero-shot, few-shot, dan chain-of-thought, dalam tugas klasifikasi sentimen teks berbahasa Indonesia. Eksperimen dilakukan pada 200 sampel data uji dari dataset SmSA (IndoNLU) menggunakan dua model open-source berukuran kecil, yaitu Qwen3-4B-Instruct-2507 dan Llama-3.2-3B-Instruct, dengan evaluasi berdasarkan akurasi, precision, recall, dan F1-score macro, dilengkapi selang kepercayaan Wilson, uji McNemar berpasangan, dan baseline kelas mayoritas. Seluruh konfigurasi mengungguli baseline kelas mayoritas (akurasi 0,460; F1 0,210) secara meyakinkan. Audit ekstraksi label menemukan bahwa implementasi awal salah mengekstraksi 18 dari 200 keluaran pada satu kondisi secara sistematis dan searah, padahal kondisi tersebut mencatat unknown rate 0%; koreksinya menaikkan akurasi kondisi tersebut dari 0,855 menjadi 0,945. Konfigurasi terbaik adalah few-shot pada Qwen3-4B-Instruct-2507 (akurasi 0,945; F1 macro 0,918). Qwen3-4B-Instruct-2507 secara signifikan mengungguli Llama-3.2-3B-Instruct pada ketiga teknik (uji McNemar eksak dengan koreksi Holm). Satu-satunya perbedaan antar teknik yang signifikan adalah keunggulan few-shot atas chain-of-thought pada Qwen3-4B-Instruct-2507 (p Holm=0,038); chain-of-thought tidak pernah menjadi teknik terbaik pada kedua model sementara biayanya puluhan kali lipat. Seluruh perbedaan antar kondisi bersumber dari kelas netral, yang ditangani kedua model jauh kurang andal dibandingkan sentimen positif dan negatif.**

Kata Kunci— prompt engineering, large language model, klasifikasi sentimen, bahasa Indonesia, zero-shot, few-shot

Submit: 05/07/2026 *Revisi: 10/07/2026* *Accepted: 14/07/2026*

I. PENDAHULUAN

Perkembangan Large Language Model (LLM) seperti GPT-4, Claude, dan model open-source lainnya telah membuka peluang baru dalam berbagai tugas pemrosesan bahasa alami (*Natural Language Processing*/NLP), termasuk klasifikasi sentimen. Berbeda dengan pendekatan tradisional yang memerlukan pelatihan model klasifikasi secara khusus (misalnya *fine-tuning* model berbasis BERT), LLM memungkinkan klasifikasi dilakukan hanya melalui instruksi berbasis teks atau yang dikenal dengan istilah *prompt*, tanpa memerlukan proses pelatihan ulang.

Meskipun pendekatan ini menawarkan efisiensi dari sisi waktu dan sumber daya, performa LLM dalam melakukan klasifikasi sangat bergantung pada bagaimana *prompt* dirancang. Berbagai teknik *prompt engineering* telah dikembangkan, seperti *zero-shot prompting* (tanpa contoh), *few-shot prompting* (dengan beberapa contoh), dan *chain-of-thought prompting* (dengan penalaran bertahap), yang masing-masing memiliki karakteristik dan tingkat efektivitas yang berbeda, tergantung pada tugas dan bahasa yang digunakan.

Dalam bahasa Indonesia, penelitian yang secara khusus membandingkan efektivitas berbagai teknik prompt engineering untuk tugas klasifikasi sentimen masih relatif terbatas. Sebagian besar studi terdahulu berfokus pada evaluasi performa LLM secara umum dibandingkan dengan model berbasis *fine-tuning* seperti IndoBERT, namun belum banyak yang mengeksplorasi secara sistematis bagaimana variasi teknik *prompt* memengaruhi hasil klasifikasi pada teks berbahasa Indonesia yang memiliki karakteristik khas, seperti campur kode (code-mixing), bahasa informal, dan penggunaan singkatan.

Berdasarkan hal tersebut, penelitian ini bertujuan untuk mengevaluasi dan membandingkan efektivitas beberapa teknik *prompt engineering* pada tugas klasifikasi sentimen teks berbahasa Indonesia, guna memberikan rekomendasi praktis bagi peneliti maupun praktisi yang ingin menerapkan LLM dalam konteks bahasa lokal.

II. TINJAUAN PUSTAKA

Penelitian mengenai klasifikasi sentimen teks berbahasa Indonesia telah berkembang dari pendekatan berbasis leksikon dan *machine learning* klasik, hingga pendekatan berbasis transformer seperti IndoBERT. Wilie dkk. [1] Memperkenalkan IndoNLU, sebuah *benchmark* yang menyediakan dataset standar untuk evaluasi tugas-tugas NLP berbahasa Indonesia, termasuk dataset SmSA untuk klasifikasi sentimen tiga kelas (positif, negatif, netral), yang banyak dijadikan acuan dalam penelitian selanjutnya.

Dengan munculnya LLM generatif, sejumlah studi mulai mengevaluasi kemampuan model tersebut dalam melakukan klasifikasi sentimen tanpa perlu melakukan *fine-tuning*. Wang dkk. [2] Mengevaluasi kemampuan ChatGPT sebagai penganalisis sentimen dan menemukan bahwa performanya secara zero-shot dapat menyaingi model yang telah dilatih secara khusus pada berbagai benchmark. Untuk konteks bahasa Indonesia secara spesifik, Nasution dkk. [3] Melakukan *benchmarking* terhadap 22 LLM open-source dibandingkan dengan ChatGPT-4 dan anotator manusia pada tugas klasifikasi sentimen dan emosi pada teks Twitter berbahasa Indonesia menggunakan pendekatan *zero-shot prompting*. Studi tersebut menemukan bahwa ChatGPT-4 mencapai performa yang sebanding dengan anotator manusia, dan beberapa model open-source berukuran besar seperti LLaMA dan DeepSeek mampu mendekati performa tersebut.

Di sisi lain, pendekatan berbasis *fine-tuning* terhadap model pralatih seperti IndoBERT masih menunjukkan performa yang kompetitif. Saputra dkk. [4] Mengembangkan IndoBERT-Sentiment, model klasifikasi sentimen berbasis konteks topik yang dilatih pada lebih dari 31.000 pasangan konteks-teks, dan menunjukkan bahwa pendekatan context-conditioned mampu mengungguli model sentimen umum berbahasa Indonesia lainnya secara signifikan. Studi lain oleh Zahra dkk. [5] Membandingkan performa metode machine learning klasik (Logistic Regression, SVM, Naive Bayes) dengan *fine-tuning* IndoBERT pada ulasan produk berbahasa Indonesia.

Studi-studi tersebut menunjukkan bahwa perbandingan antarmodel dan antar LLM dengan pendekatan *fine-tuning* untuk teks berbahasa Indonesia telah cukup banyak dilakukan. Nasution dkk. [3], misalnya, telah membenchmark 22 LLM open-source, namun seluruhnya pada satu teknik *prompt* yang sama (*zero-shot*). Celah yang belum banyak dieksplorasi adalah arah sebaliknya, yaitu membandingkan beberapa teknik prompt engineering dalam model yang sama untuk teks berbahasa Indonesia, serta memeriksa apakah teknik yang efektif pada satu model tetap efektif pada model lain. Kontribusi penelitian ini karenanya ada tiga: (1) perbandingan tiga teknik prompt (*zero-shot, few-shot, chain-of-thought*) di dalam model yang sama pada dataset SmSA; (2) pelaporan biaya komputasi tiap teknik, yang jarang disertakan dalam studi sejenis, padahal menentukan kelayakan penerapan pada perangkat lokal; dan (3) pelaporan hasil disertai uji statistik dan baseline, sehingga selisih yang bermakna dapat dibedakan dari variasi acak pada ukuran sampel yang terbatas.

III. METODE

A. Dataset

Penelitian ini menggunakan dataset SmSA (Sentiment Analysis), salah satu subset dari benchmark IndoNLU, yang diakses melalui pustaka Hugging Face Datasets (indonlp/indonlu, konfigurasi smsa). Dataset ini terdiri atas teks berbahasa Indonesia yang telah diberi label sentimen ke dalam tiga kelas: positif, negatif, dan netral. Nama kelas diambil langsung dari metadata fitur dataset *(features['label'].names)* untuk menghindari kesalahan pemetaan urutan label.

Evaluasi dilakukan pada split test resmi dari benchmark IndoNLU, bukan pada data latih, untuk menjaga validitas hasil. Karena keterbatasan sumber daya komputasi (inferensi dilakukan secara lokal), digunakan subset acak sebanyak 200 data dari split test (random_state=42 agar hasil dapat direproduksi). Seluruh split test dapat digunakan apabila sumber daya komputasi memungkinkan, dengan

mengubah parameter jumlah sampel dalam skrip eksperimen.

Pengambilan subset dilakukan secara acak sederhana tanpa stratifikasi, sehingga distribusi kelas pada subset yang digunakan tidak seimbang, yaitu 92 sampel negatif, 73 sampel positif, dan 35 sampel netral. Ketimpangan ini perlu diperhatikan dalam menafsirkan metrik per kelas, khususnya pada kelas netral yang memiliki jumlah sampel paling sedikit. Sebagai contoh, pada teknik few-shot prompting, digunakan tiga contoh teks beserta labelnya yang disusun secara manual dan terpisah dari data uji, untuk menghindari kebocoran data (*data leakage*) antara contoh prompt dan data yang dievaluasi. Perlu dicatat bahwa hanya satu set contoh yang digunakan, tanpa variasi jumlah maupun pemilihan contoh.

B. Model LLM yang Diuji

Eksperimen dilakukan menggunakan dua model large language model (LLM) open-source berukuran kecil yang dijalankan secara lokal, yaitu:

a) *Qwen3-4B-Instruct-2507*: model instruction-tuned dari kelompok Qwen3 (Alibaba) dengan 4 miliar parameter.
b) *Llama-3.2-3B-Instruct*: model instruction-tuned dari kelompok Llama 3.2 (Meta) dengan 3 miliar parameter, kuantisasi Q8.

Kedua model dipilih karena berukuran cukup kecil untuk dijalankan sepenuhnya pada perangkat lokal dengan sumber daya GPU terbatas, sehingga eksperimen dapat dilakukan tanpa bergantung pada layanan API berbayar. Pemilihan dua model dari kelompok arsitektur yang berbeda juga memungkinkan pengujian apakah efektivitas suatu teknik prompt bersifat umum atau bergantung pada model yang digunakan.

C. Teknik Prompt yang Dievaluasi

Penelitian ini mengevaluasi tiga teknik prompt engineering berikut:

a) *Zero-shot prompting*: model diminta melakukan klasifikasi sentimen tanpa diberikan contoh sebelumnya, hanya instruksi langsung mengenai kategori yang tersedia (positif, negatif, netral), dengan permintaan agar model menjawab hanya dengan satu label kata.
b) *Few-shot prompting*: model diberikan tiga contoh teks beserta label sentimennya sebelum diminta untuk mengklasifikasikan teks baru.
c) *Chain-of-thought prompting*: model diminta menuliskan penalaran singkat terlebih dahulu, kemudian memberikan jawaban akhir dalam format baku "Jawaban akhir: <label>" pada baris terakhir, untuk memudahkan proses ekstraksi label secara otomatis.

Ketiga teknik diimplementasikan sebagai template *prompt* tetap agar perbandingan antar teknik konsisten dan dapat direproduksi. Setiap teknik hanya diwakili oleh satu formulasi *prompt* dan tanpa variasi parafrase serta terdapat implikasi keterbatasan.

D. Parameter Inferensi

Setiap pemanggilan model dilakukan dengan parameter temperature = 0,0 agar keluaran model bersifat deterministik, sehingga perbandingan antar teknik *prompt* tidak dipengaruhi oleh variasi acak (*randomness*) pada proses *decoding*. Panjang konteks (num_ctx) diatur sebesar 2048 token. Apabila terjadi kegagalan pemanggilan (misalnya *timeout*), sistem melakukan percobaan ulang secara otomatis hingga tiga kali sebelum mencatat keluaran kosong untuk sampel tersebut.

E. Ekstraksi Label dari Keluaran Model

Karena LLM generatif dapat menghasilkan keluaran berupa teks bebas (bukan label terstruktur), setiap keluaran mentah diproses melalui tahap ekstraksi label berbasis pencocokan pola (regular expression). Untuk teknik *chain-of-thought*, sistem terlebih dahulu mencari pola "Jawaban akhir: <label>"; apabila pola tersebut tidak ditemukan, sistem menggunakan mekanisme cadangan (*fallback*) berupa pencarian kata kunci label pada seluruh teks keluaran. Apabila tidak ada label valid yang dapat diekstraksi, prediksi dicatat sebagai "*unknown*" dan tetap dimasukkan ke dalam perhitungan metrik sebagai bentuk pelaporan yang jujur terhadap keterbatasan format keluaran model. Karena seluruh metrik bergantung pada ketepatan tahap ini, setiap keluaran mentah diaudit dengan mencatat jalur ekstraksi yang digunakan (jangkar atau cadangan), jumlah label berbeda yang muncul dalam teks, serta kesesuaian antara label awal dan hasil ekstraksi ulang yang memilih label berdasarkan posisi kemunculannya, bukan urutan prioritas nama label.

F. Metrik Evaluasi dan Uji Statistik

Evaluasi dilakukan menggunakan metrik akurasi (*accuracy*), precision, recall, dan F1-score dengan perhitungan *macro-average* agar performa pada setiap kelas (positif, negatif, netral) diperhitungkan secara seimbang, menggunakan pustaka scikit-learn. Proporsi kegagalan ekstraksi label (unknown rate) dilaporkan sebagai indikator tambahan untuk mengukur konsistensi format keluaran.

Karena ukuran sampel yang relatif kecil (n=200), pelaporan nilai metrik saja tidak memadai untuk menilai apakah selisih antar kondisi bermakna. Oleh karena itu, penelitian ini juga melaporkan: (a) selang kepercayaan Wilson 95% untuk akurasi setiap kondisi; (b) selang kepercayaan bootstrap 95% (5.000 resampel) untuk F1-score macro pada kondisi yang datanya tersedia pada tingkat prediksi per-sampel; (c) uji McNemar berpasangan untuk membandingkan dua kondisi yang dievaluasi pada sampel yang sama; serta (d) baseline kelas mayoritas (selalu memprediksi kelas terbanyak) sebagai pembanding dasar. Sembilan perbandingan berpasangan dilakukan dan dikelompokkan ke dalam dua kelompok hipotesis: perbandingan antar teknik prompt di dalam model yang sama, dan perbandingan antarmodel pada teknik yang sama, dengan koreksi Holm yang diterapkan di dalam masing-

masing kelompok. Perbedaan yang selang kepercayaannya tumpang tindih dilaporkan sebagai tidak dapat dibedakan, bukan sebagai perbedaan kinerja.

G. *Prosedur Eksperimen*

Prosedur eksperimen dilakukan melalui langkah-langkah berikut:

a) Menyiapkan lingkungan inferensi lokal menggunakan Ollama dan mengunduh (*pull*) model yang akan diuji.
b) Memuat subset data uji SmSA sesuai dengan konfigurasi.
c) Untuk setiap kombinasi model dan teknik prompt, membangun prompt sesuai template, mengirimkannya ke model melalui Ollama, kemudian mengekstraksi label prediksi dari keluaran mentah sesuai prosedur.
d) Menyimpan seluruh prediksi beserta teks asli, label sebenarnya (*ground truth*), dan keluaran mentah model untuk keperluan analisis kesalahan (*error analysis*).
e) Menghitung metrik evaluasi untuk setiap kombinasi model dan teknik prompt, kemudian menyimpannya sebagai dasar penyusunan tabel hasil.

IV. HASIL DAN PEMBAHASAN

A. *Audit Ekstraksi Label*

Karena seluruh metrik pada penelitian ini bergantung pada ketepatan konversi keluaran teks bebas menjadi label, audit terhadap proses ekstraksi dilakukan terlebih dahulu sebelum hasil klasifikasi dianalisis. Setiap keluaran mentah diperiksa ulang untuk mencatat jalur ekstraksi yang digunakan (pola jangkar "Jawaban akhir:" atau mekanisme cadangan berbasis kata kunci), jumlah label yang berbeda dalam teks keluaran, serta kesesuaian antara label hasil ekstraksi awal dan hasil ekstraksi ulang. Rekapitulasinya disajikan pada Tabel I.

TABEL I. AUDIT EKSTRAKSI LABEL (N=200 PER KONDISI)

Model	Teknik	Jangkar	Cadangan	Unknown	Multi-label	Koreksi
Qwen3-4B	Zero-shot	0	200	0	0	0
	Few-shot	0	200	0	23	18
	CoT	200	0	0	192	0
Llama-3.2-3B	Zero-shot	0	200	0	0	0
	Few-shot	0	192	8	55	0
	CoT	200	0	0	131	0

Audit ini menghasilkan tiga temuan yang memengaruhi penafsiran seluruh hasil selanjutnya.

Pertama, kekhawatiran bahwa mekanisme cadangan dapat merusak label pada teknik *chain-of-thought* ternyata tidak terbukti. Pola jangkar berhasil ditemukan pada seluruh 200 keluaran *chain-of-thought* untuk kedua model, sehingga mekanisme cadangan sama sekali tidak digunakan dalam teknik ini dan tidak ditemukan ketidaksesuaian pada ekstraksi ulang. Meskipun demikian, audit juga menunjukkan bahwa rancangan jangkar tersebut memang esensial: 192 dari 200 keluaran *chain-of-thought* Qwen3-4B-Instruct-2507 dan 131 dari 200 keluaran Llama-3.2-3B-Instruct memuat lebih dari satu nama label dalam teksnya, sehingga pencarian kata kunci tanpa jangkar hampir pasti akan menghasilkan label yang keliru.

Kedua, dan justru di luar dugaan, kesalahan ekstraksi yang sesungguhnya ditemukan pada teknik few-shot Qwen3-4B-Instruct-2507. Sebanyak 18 dari 200 keluaran diekstraksi secara keliru oleh implementasi awal. Seluruh 18 kasus memiliki pola yang identik: model menuliskan kesimpulannya pada baris pertama (misalnya "Sentimen: netral"), lalu menyusul dengan penjelasan yang menyebutkan label lain, seperti "tanpa ekspresi emosi yang jelas positif atau negatif". Implementasi awal memilih label berdasarkan urutan prioritas nama label alih-alih posisi kemunculannya di dalam teks, sehingga kata "positif" pada bagian penjelasan mengalahkan label "netral" yang merupakan jawaban yang sebenarnya. Akibatnya bersifat sistematis dan searah: seluruh 18 kasus keliru diberi label positif, padahal jawaban model adalah netral, dan seluruhnya menjadi benar setelah dikoreksi. Koreksi ini meningkatkan akurasi kondisi tersebut dari 0,855 menjadi 0,945.

Ketiga, teknik few-shot pada Llama-3.2-3B-Instruct merupakan satu-satunya kondisi dengan *unknown rate* yang bukan nol, yaitu 8 sampel (4%) yang keluarannya tidak memuat label valid sama sekali. Kedelapan sampel tersebut tetap disertakan dalam perhitungan metrik sebagai prediksi yang salah.

Seluruh analisis menggunakan label hasil ekstraksi ulang yang telah diaudit. Temuan ini sekaligus menggarisbawahi bahwa *unknown rate* sebesar 0% bukan indikator yang memadai untuk menilai keandalan ekstraksi: kondisi few-shot Qwen3-4B-Instruct-2507 mencatat *unknown rate* 0%, namun memuat 18 label yang keliru.

B. *Performa Klasifikasi*

Hasil evaluasi setelah koreksi ekstraksi label disajikan pada Tabel II, lengkap dengan selang kepercayaan Wilson 95% untuk akurasi dan selang kepercayaan bootstrap 95% (5.000 resampel) untuk F1-score macro.

TABEL II. HASIL EVALUASI SETELAH KOREKSI EKSTRAKSI LABEL (N=200; CI WILSON UNTUK AKURASI, BOOTSTRAP UNTUK F1)

Model	Teknik	Akurasi (95% CI)	F1 macro (95% CI)	Prec. / Recall
Baseline	—	0,460 [0,392–0,529]	0,210	0,153 / 0,333
Qwen3-4B	Zero-shot	0,910 [0,862–0,942]	0,857 [0,794–0,912]	0,907 / 0,839
	Few-shot	0,945 [0,904–0,969]	0,918 [0,867–0,960]	0,948 / 0,900
	CoT	0,895 [0,845–0,930]	0,824 [0,750–0,886]	0,923 / 0,805
Llama-3.2-3B	Zero-shot	0,785 [0,723–0,836]	0,645 [0,573–0,717]	0,802 / 0,665
	Few-shot	0,810 [0,750–0,858]	0,775 [0,706–0,836]	0,799 / 0,758
	CoT	0,840 [0,783–0,884]	0,783 [0,717–0,843]	0,782 / 0,786

Seluruh kondisi mengungguli baseline kelas mayoritas (akurasi 0,460; F1 0,210) dengan selisih yang signifikan. Konfigurasi terbaik adalah *few-shot prompting* pada Qwen3-4B-Instruct-2507 dengan akurasi 0,945 dan F1-

score 0,918, sedangkan konfigurasi terlemah adalah *zero-shot* pada Llama-3.2-3B-Instruct (akurasi 0,785; F1-score 0,645). Perlu dicatat bahwa satu-satunya kondisi dengan *unknown rate* yang bukan nol adalah *few-shot* pada Llama-3.2-3B-Instruct (4%), yang tetap mencatat F1-score lebih tinggi daripada *zero-shot* pada model yang sama.

C. *Uji Signifikansi Statistik*

Seluruh kondisi dievaluasi pada 200 sampel yang sama, sehingga perbandingan antar kondisi dilakukan menggunakan uji McNemar eksak pada pasangan prediksi per-sampel. Sembilan perbandingan diuji dan dikelompokkan ke dalam dua kelompok hipotesis: kelompok A membandingkan teknik *prompt* dalam model yang sama (enam perbandingan), dan kelompok B membandingkan kedua model dengan teknik yang sama (tiga perbandingan). Koreksi Holm diterapkan pada masing-masing kelompok. Hasilnya disajikan pada Tabel III.

TABEL III. UJI MCNEMAR EKSAK BERPASANGAN (N=200; KOREKSI HOLM DI DALAM MASING-MASING KELOMPOK)

Kel.	Perbandingan	Diskordan	p	p (Holm)	Signif.
A	Qwen3: zero-shot vs few-shot	9	0,039	0,195	tidak
	Qwen3: zero-shot vs CoT	5	0,375	1,000	tidak
	Qwen3: few-shot vs CoT	12	0,006	0,038	ya
	Llama: zero-shot vs few-shot	27	0,442	1,000	tidak
	Llama: zero-shot vs CoT	29	0,061	0,246	tidak
	Llama: few-shot vs CoT	34	0,392	1,000	tidak
B	Qwen3 vs Llama (zero-shot)	27	<0,001	<0,001	ya
	Qwen3 vs Llama (few-shot)	35	<0,001	<0,001	ya
	Qwen3 vs Llama (CoT)	25	0,043	0,043	ya

Pada kelompok B, ketiga perbandingan antar model signifikan secara statistik: Qwen3-4B-Instruct-2507 mengungguli Llama-3.2-3B-Instruct pada zero-shot dan few-shot dengan bukti yang sangat kuat (keduanya $p<0,001$ setelah koreksi), serta pada chain-of-thought dengan bukti yang jauh lebih tipis ($p=0,043$). Menyempitnya bukti pada kondisi chain-of-thought konsisten dengan berkurangnya selisih F1-score antar model pada kondisi tersebut.

Pada kelompok A, hanya satu perbandingan antar teknik yang bertahan setelah koreksi, yaitu few-shot mengungguli chain-of-thought pada Qwen3-4B-Instruct-2507 (12 pasangan diskordan; $p=0,006$; p Holm=0,038). Ini merupakan satu-satunya bukti dalam penelitian ini bahwa pemilihan teknik prompt berpengaruh nyata terhadap kinerja.

Perbandingan antar teknik lainnya tidak mencapai signifikansi. Pada Qwen3-4B-Instruct-2507, *zero-shot dan chain-of-thought* menghasilkan prediksi identik pada 195 dari 200 sampel ($p=0,375$), sementara keunggulan few-shot atas zero-shot hanya bersandar pada 9 pasangan diskordan dan tidak bertahan setelah koreksi (p Holm=0,195). Pada Llama-3.2-3B-Instruct, tidak satu pun dari ketiga perbandingan teknik mencapai signifikansi, termasuk keunggulan *chain-of-thought* atas *zero-shot* yang secara nilai metrik tampak besar (F1 0,783 berbanding 0,645), namun hanya menghasilkan $p=0,061$ sebelum koreksi dan $p=0,246$ setelahnya.

Secara ringkas, perbedaan antar model jauh lebih kokoh secara statistik dibandingkan perbedaan antar teknik prompt. Pada kedua model, *chain-of-thought* tidak pernah menjadi teknik terbaik, dan pada model yang lebih kuat, teknik tersebut terbukti lebih buruk daripada *few-shot*.

D. *Analisis per Kelas*

Seluruh perbedaan antar kondisi bersumber dari satu kelas. Pada kelas positif, kedua model hampir sempurna pada seluruh kondisi (69–72 benar dari 73). Pada kelas negatif, Qwen3-4B-Instruct-2507 mencatat 91–92 benar dari 92, sedangkan Llama-3.2-3B-Instruct mencatat 75–81 benar dari 92. Kelas netral, yang hanya diwakili oleh 35 sampel, merupakan sumber utama seluruh kesalahan sebagaimana dirinci pada Tabel IV.

TABEL IV. RINCIAN KLASIFIKASI KELAS NETRAL (N=35)

Model	Teknik	Benar	→ positif	→ negatif	Unknown
Qwen3-4B	Zero-shot	19	9	7	0
	Few-shot	25	5	5	0
	CoT	15	12	8	0
Llama-3.2-3B	Zero-shot	5	20	10	0
	Few-shot	18	8	7	2
	CoT	19	10	6	0

Tabel IV memperlihatkan bahwa setiap perubahan performa yang teramati pada Tabel II dapat ditelusuri ke penanganan kelas netral. Pada Qwen3-4B-Instruct-2507, few-shot merupakan kondisi terbaik untuk kelas ini (25 dari 35 benar), sedangkan chain-of-thought merupakan yang terburuk (15 dari 35), yang menjelaskan mengapa satu-satunya perbedaan teknik yang signifikan dalam penelitian ini adalah perbandingan antara kedua kondisi tersebut.

Pada Llama-3.2-3B-Instruct, kondisi *zero-shot* menunjukkan kegagalan yang mencolok pada kelas netral, yaitu hanya 5 dari 35 sampel yang benar, dengan 20 di antaranya dialihkan ke kelas positif. Baik *few-shot* (18 benar) maupun *chain-of-thought* (19 benar) memperbaiki kondisi ini secara substansial. Meskipun demikian, sebagaimana ditunjukkan pada Tabel III, perbaikan tersebut tidak mencapai signifikansi statistik karena disertai peningkatan kesalahan pada kelas negatif, sehingga keuntungan bersihnya pada tingkat sampel secara keseluruhan menjadi terbatas.

Kelas netral hanya diwakili 35 sampel karena pengambilan sampel dilakukan secara acak tanpa stratifikasi. Seluruh angka pada Tabel IV karenanya memiliki ketidakpastian yang besar, dan simpulan mengenai kelas ini sebaiknya dikonfirmasi melalui pengambilan sampel berstrata.

E. *Analisis Efisiensi Komputasi*

Waktu inferensi per sampel dicatat sebagai indikator biaya komputasi. Karena distribusi latensi memiliki ekor kanan yang panjang akibat cold start dan variasi panjang keluaran, Tabel V melaporkan median dan rata-rata.

TABEL V. WAKTU INFERENSI PER SAMPEL (DETIK)

Model	Teknik	Median	Rata-rata	Total
Qwen3-4B	Zero-shot	0,57	0,69	137,7
	Few-shot	2,18	7,30	1.461,0
	CoT	38,41	39,97	7.993,5
Llama-3.2-3B	Zero-shot	1,01	1,26	251,6
	Few-shot	23,03	24,60	4.920,5
	CoT	20,89	22,11	4.422,9

Zero-shot merupakan teknik termurah pada kedua model, dengan median di bawah 1,1 detik per sampel. *Chain-of-thought* merupakan yang termahal pada Qwen3-4B-Instruct-2507, yaitu 38,41 detik atau sekitar 67 kali dari median *zero-shot*.

Dua pengamatan pada Tabel V memerlukan kehati-hatian. ***Pertama***, latensi few-shot pada Llama-3.2-3B-Instruct (median 23,03 detik) justru lebih tinggi daripada *chain-of-thought* pada model yang sama (20,89 detik). Hal ini berlawanan dengan ekspektasi, karena *chain-of-thought* menghasilkan keluaran yang jauh lebih panjang sehingga seharusnya lebih lambat. ***Kedua***, pada Qwen3-4B-Instruct-2507 rata-rata latensi *few-shot* (7,30 detik) jauh melampaui mediannya (2,18 detik), yang menandakan adanya sejumlah kecil pemanggilan yang sangat lambat. Kedua anomali tersebut lebih mungkin mencerminkan perbedaan kondisi eksekusi (misalnya beban sistem atau konfigurasi runtime pada sesi pengukuran) daripada karakteristik intrinsik teknik *few-shot*. Angka latensi *few-shot* pada Tabel V karenanya sebaiknya diperlakukan sebagai indikatif, dan pengukuran ulang dalam satu sesi terkendali diperlukan sebelum perbandingan biaya antar teknik dapat dianggap final.

Terlepas dari catatan tersebut, satu simpulan tetap kokoh karena selisihnya berada pada orde besaran yang sama. Pada Qwen3-4B-Instruct-2507, *chain-of-thought* terbukti secara statistik lebih buruk daripada few-shot (Tabel III) sekaligus sekitar 18 kali lebih lambat berdasarkan median. Dengan demikian, pada model tersebut, *chain-of-thought* tidak memiliki justifikasi, baik dari sisi performa maupun biaya. Pada Llama-3.2-3B-Instruct, *chain-of-thought* memang mencatat F1-score tertinggi, namun keunggulannya tidak terbukti signifikan sementara biayanya tetap tinggi; sebagai perbandingan, konfigurasi *few-shot* pada model yang lebih kuat mencapai F1-score 0,918 dengan median latensi 2,18 detik, jauh mengungguli F1-score 0,783 pada 20,89 detik.

F. Analisis Kesalahan (Error Analysis)

Sebagaimana telah diteliti, kesalahan kedua model hampir seluruhnya terkonsentrasi pada kelas netral. Pemeriksaan terhadap kalimat yang salah diklasifikasikan memperlihatkan dua pola yang berulang.

Pola pertama adalah kalimat netral yang bersifat informatif atau faktual, namun memuat kata berkonotasi positif, sehingga diklasifikasikan sebagai kalimat positif. Contohnya: “Salah satu kegemaran anak remaja Indonesia sekarang adalah mendengarkan K-pop.” dan “Kemarin aku setelah nyobain kopinya Warung Kopi Nako.”. Kedua kalimat tidak menyatakan penilaian, tetapi kata seperti “kegemaran” dan “nyobain” tampaknya cukup untuk memicu prediksi positif. Pola ini merupakan sumber kesalahan terbesar, mencapai 20 dari 35 sampel netral pada kondisi *zero-shot* Llama-3.2-3B-Instruct.

Pola kedua adalah kalimat yang menyampaikan kritik atau kekecewaan secara implisit tanpa kata negatif yang kuat, sehingga sentimen negatifnya tidak tertangkap dan kalimat tersebut dinilai netral. Contohnya “Kadang bingung sih kenapa mulut netijen pada tidak bisa dijaga, bisa nya behina terus. aneh.”.

Kedua pola menunjukkan bahwa kesulitan utama bukanlah membedakan polaritas positif dari negatif (kedua model hampir tidak pernah keliru pada dimensi tersebut) melainkan membedakan pernyataan faktual dari ungkapan sentimen. Karakteristik teks berbahasa Indonesia informal pada dataset ini, yang banyak mengandung campur kode dan singkatan, kemungkinan akan memperumit pembedaan tersebut.

Ditemukan dalam proses audit memiliki pola yang serupa. Ke-18 keluaran yang keliru diekstraksi seluruhnya merupakan kasus di mana model menjawab “netral” dengan tepat, namun penjelasan yang menyertainya menyebut kata “positif” dalam konteks penyangkalan. Dengan kata lain, kelas netral tidak hanya sulit bagi model, tetapi juga rawan terhadap proses ekstraksi label, karena penalaran mengenai ketiadaan sentimen secara alami memerlukan penyebutan nama-nama sentimen yang tidak berlaku.

V. KESIMPULAN

Penelitian ini mengevaluasi tiga teknik *prompt engineering*: *zero-shot*, *few-shot*, dan *chain-of-thought* untuk klasifikasi sentimen teks berbahasa Indonesia pada dua model LLM open-source berukuran kecil, yaitu Qwen3-4B-Instruct-2507 dan Llama-3.2-3B-Instruct, menggunakan 200 sampel data uji dari dataset SmSA (IndoNLU), dengan audit ekstraksi label, uji signifikansi berpasangan, dan koreksi perbandingan berganda.

Temuan yang didukung oleh bukti statistik adalah sebagai berikut. Pertama, konfigurasi terbaik adalah *few-shot prompting* pada Qwen3-4B-Instruct-2507 dengan akurasi 0,945 dan F1-score macro 0,918. Kedua, Qwen3-4B-Instruct-2507 secara signifikan mengungguli Llama-3.2-3B-Instruct pada ketiga teknik *prompt* tersebut. Ketiga, satu-satunya perbedaan antar teknik *prompt* yang terbukti signifikan adalah keunggulan *few-shot* atas *chain-of-thought* pada Qwen3-4B-Instruct-2507 (p Holm=0,038); *chain-of-thought* tidak pernah menjadi teknik terbaik pada kedua model, sementara biayanya mencapai puluhan kali lipat teknik lainnya. Keempat, seluruh perbedaan antar kondisi bersumber dari kelas netral, sedangkan kelas positif dan negatif ditangani dengan baik pada semua kondisi.

Kontribusi metodologis penelitian ini terletak pada audit ekstraksi label. Audit tersebut menemukan bahwa implementasi awal salah mengekstraksi 18 dari 200 keluaran pada satu kondisi secara sistematis dan searah, sehingga akurasi kondisi tersebut dilaporkan 9 poin persentase lebih rendah dari yang seharusnya. Yang perlu digarisbawahi, kondisi tersebut mencatat *unknown rate* sebesar 0%, sehingga kesalahan ini tidak akan terdeteksi apabila keandalan ekstraksi hanya dinilai berdasarkan *unknown rate*. Sebaliknya, kekhawatiran awal bahwa teknik *chain-of-thought* paling rawan terhadap kesalahan ekstraksi

tidak terbukti, karena pola jangkar berhasil digunakan pada seluruh keluarannya. Bagi penelitian sejenis yang menggunakan LLM generatif sebagai pengklasifikasi, pelaporan jalur ekstraksi dan verifikasi ulang label sebaiknya diperlakukan sebagai bagian baku dari prosedur, bukan sebagai pemeriksaan tambahan.

Beberapa hal tidak dapat disimpulkan dari data ini. Perbedaan antar teknik prompt pada Llama-3.2-3B-Instruct tidak satu pun mencapai signifikansi, termasuk keunggulan *chain-of-thought* yang secara nilai metrik tampak besar. Demikian pula, pola bahwa *chain-of-thought* tampak lebih bermanfaat bagi model yang lebih lemah tidak dapat diatribusikan pada kapabilitas model, karena kedua model berbeda dalam arsitektur, jumlah parameter, dan tingkat kuantisasi. Perbandingan biaya yang melibatkan teknik *few-shot* juga masih terkendala oleh anomali pengukuran latensi.

Keterbatasan penelitian ini meliputi: (1) ukuran sampel 200 memberikan daya uji yang terbatas untuk mendeteksi selisih kecil antar teknik; (2) pengambilan sampel tidak berstrata sehingga kelas netral, yang menjadi sumber seluruh kesalahan, hanya terwakili 35 sampel; (3) hanya dua model yang dibandingkan dengan perbedaan arsitektur, ukuran, dan kuantisasi yang saling terkonfound; (4) setiap teknik hanya diwakili satu formulasi prompt tanpa variasi parafrase; (5) tidak ada baseline model terlatih seperti IndoBERT pada subset yang sama; serta (6) pengukuran latensi *few-shot* berasal dari kondisi eksekusi yang tampaknya berbeda sehingga belum sepenuhnya dapat dibandingkan.

Penelitian selanjutnya disarankan untuk memperluas evaluasi ke seluruh split test SmSA dengan pengambilan sampel berstrata agar daya uji memadai dan kelas netral terwakili proporsional; menguji minimal empat hingga enam model, idealnya beberapa ukuran dari kelompok arsitektur yang sama sehingga kapabilitas dapat divariasikan secara terkendali; menjalankan tiga atau lebih parafrase *prompt* per teknik dan melaporkan rata-rata beserta simpangan bakunya; menyertakan baseline IndoBERT yang di-fine-tune pada subset uji yang sama; mengukur ulang seluruh latensi dalam satu sesi terkendali sambil mencatat jumlah token keluaran sebagai metrik biaya yang portabel; serta menerapkan audit ekstraksi label sebagai prosedur baku sejak awal eksperimen.

DAFTAR PUSTAKA

[1] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar, and A. Purwarianti, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," *Proc. 1st Conf. Asia-Pacific Chapter Assoc. Comput. Linguistics 10th Int. Joint Conf. Natural Lang. Process.*, pp. 843–857, 2020.

[2] Z. Wang, Q. Xie, Y. Feng, Z. Ding, Z. Yang, and R. Xia, "Is ChatGPT a good sentiment analyzer? A preliminary study," in *Proc. 1st Conf. Language Modeling (COLM)*, 2024. [Online]. Available: arXiv:2304.04339

[3] A. H. Nasution, A. Onan, Y. Murakami, W. Monika, and A. Hanafiah, "Benchmarking open-source large language models for sentiment and emotion classification in Indonesian tweets," *IEEE Access*, vol. 13, pp. 94009–94025, 2025, doi: 10.1109/ACCESS.2025.3574629.

[4] M. A. A. Saputra, A. Alamsyah, D. P. Ramadhani, T. S. Siadari, and H. Fakhrurroja, "IndoBERT-Sentiment: Context-conditioned sentiment classification for Indonesian text," *arXiv preprint arXiv:2604.07057*, 2026, doi: 10.48550/arXiv.2604.07057.

[5] N. Z. Zahra, S. Farhanatussaidah, N. N. Afifah, L. Muthoharoh, A. Satria, and M. C. T. Manullang, "Benchmarking logistic regression, SVM, Naive Bayes, and IndoBERT fine-tuning for sentiment analysis on Indonesian product reviews," *arXiv preprint arXiv:2605.03439*, 2026, doi: 10.48550/arXiv.2605.03439.