BAB II.
TINJAUAN PUSTAKA

2.1. Studi Sebelumnya

Meshkin, dkk. melakukan penelitian empiris untuk membandingkan kemampuan *Zero-shot* dan *Few-shot Prompting* dalam domain-spesifik, khususnya pada tugas ekstraksi dan klasifikasi informasi biomedis. Penelitian ini diterbitkan di jurnal *Briefings in Bioinformatics (Q1, Oxford University Press)* dan mengevaluasi berbagai model LLM pada dataset klinis yang memerlukan pemahaman terminologi teknis yang mendalam. Hasil penelitian menunjukkan bahwa *Few-shot Prompting* secara konsisten mengungguli *Zero-shot* pada seluruh tugas domain-spesifik yang diuji, dengan peningkatan F1-score rata-rata sebesar 8–15% pada tugas ekstraksi entitas dan klasifikasi klinis. Penelitian ini juga menegaskan bahwa contoh yang diberikan dalam *few-shot prompt* berfungsi sebagai panduan konteks yang membantu model memahami pola jawaban yang tepat sesuai karakteristik domain, sementara *Zero-shot* cenderung menghasilkan respon yang terlalu umum dan kurang tepat sasaran untuk pertanyaan-pertanyaan yang memerlukan pengetahuan teknis spesifik. Temuan ini relevan langsung dengan penelitian ini, di mana *chatbot* PT. Samudra Teknik Sejahtera juga beroperasi pada domain teknis yang spesifik dengan ratusan SKU produk dan kebijakan perusahaan yang kompleks, sehingga pola keunggulan *Few-shot* di domain teknis yang ditemukan Meshkin, dkk. memberikan landasan teoretis yang kuat bagi hasil penelitian ini[14].

Sivarajkumar, dkk. melakukan evaluasi empiris terhadap berbagai strategi *prompt engineering* untuk ekstraksi informasi klinis menggunakan pretrained language models dalam konteks *zero-shot* dan *few-shot*. Penelitian ini mengevaluasi berbagai jenis *prompt* (simple prefix, simple cloze, chain of thought, anticipatory, heuristic, dan ensemble) pada 5 tugas ekstraksi informasi klinis menggunakan GPT-3.5, Gemini, dan LLaMA-2. Hasil penelitian menunjukkan bahwa tailoring *prompt* yang spesifik untuk setiap tugas sangat penting untuk performa tinggi LLM dalam *zero-shot* clinical NLP. GPT-3.5

mencapai akurasi 0,96 dengan heuristic *prompts* dan 0,94 dalam biomedical evidence extraction. Heuristic *prompts*, bersama dengan chain of thought *prompts*, terbukti sangat tepat di berbagai tugas. *Few-shot prompting* meningkatkan performa dalam skenario kompleks, dan pendekatan ensemble memanfaatkan kekuatan dari multiple *prompts*. GPT-3.5 secara konsisten mengungguli Gemini dan LLaMA-2 di semua tugas dan jenis *prompt*. Penelitian ini memberikan kerangka evaluasi yang komprehensif untuk membandingkan berbagai strategi *prompting*, termasuk *Zero-shot* dan *Few-shot*, yang dapat diadaptasi untuk konteks *customer service* dalam mengukur akurasi, konsistensi, dan robustness respon *chatbot*[15].

Reynolds dan McDonell memperkenalkan konsep *prompt programming* sebagai metodologi untuk mengarahkan perilaku *Large Language Models* (LLM) melalui desain instruksi yang strategis tanpa memerlukan modifikasi parameter internal model. Penelitian ini mendemonstrasikan bahwa efektivitas model lebih ditentukan oleh kualitas instruksi utama (*meta-instruction*) daripada sekadar jumlah contoh yang diberikan, di mana instruksi yang dirancang dengan sangat jelas sering kali mampu menyamai atau melampaui performa metode *few-shot* tradisional. Dengan memandang LLM sebagai simulator teks, mereka berpendapat bahwa *prompt* berfungsi sebagai kondisi awal yang menentukan jalur penalaran model, sebuah temuan yang mendukung argumen bahwa desain *prompt* yang presisi dapat meminimalkan ketergantungan pada *prompt engineering* yang rumit tanpa menurunkan performa. Hasil ini memberikan landasan teori yang krusial bagi pengembangan sistem berbasis AI, seperti dalam layanan pelanggan, yang menunjukkan bahwa melalui teknik rekayasa informasi yang tepat, metode *zero-shot* maupun *few-shot* dapat dioptimalkan untuk menghasilkan respon yang logis dan akurat pada tugas-tugas yang kompleks [7].

Haromain, dkk. mengembangkan konsep *prompt engineering* sebagai pendekatan sistematis untuk mengoptimalkan *Large Language Model*s (LLM) dalam tugas retrieval informasi domain-spesifik seperti obat dan vitamin, dengan membagi strategi menjadi *prompt* manual (*zero-shot* bebas dan *few-shot*

terbimbing) serta integrasi *Retrieval-Augmented Generation (RAG)* untuk mengurangi halusinasi melalui pencarian vektor eksternal. Penelitian ini menunjukkan bahwa *few-shot prompting*, yang menyertakan contoh input-output terstruktur, secara signifikan melampaui *zero-shot* dalam metrik evaluasi seperti BERTScore F1 (71.64% vs 70.30%) dan ROUGE, karena contoh-contoh tersebut berfungsi sebagai panduan konteks yang memperkaya pemahaman model terhadap pola respon akurat. Dengan memanfaatkan RAG yang melibatkan embedding query, pencarian kemiripan cosine, dan augmentasi *prompt* mereka membuktikan peningkatan akurasi 2-5% pada kombinasi *few-shot*+RAG, menjadikannya fondasi teori esensial bagi aplikasi *chatbot* layanan pelanggan yang memerlukan respon presisi tinggi tanpa *Fine-tuning* berat[16].

Brown dkk. memperkenalkan GPT-3, sebuah *autoregressive language model* dengan 175 miliar parameter yang dievaluasi secara komprehensif dalam tiga kondisi *prompting*: *zero-shot* (tanpa contoh), *one-shot* (satu contoh), dan *few-shot* (beberapa contoh) pada lebih dari dua puluh *benchmark* NLP. Penelitian ini membuktikan bahwa *few-shot prompting* secara konsisten mengungguli *zero-shot* di seluruh dataset, di mana performa *few-shot* meningkat lebih cepat seiring skala model dibandingkan *zero-shot*. Sebagai contoh, pada TriviaQA, GPT-3 mencapai akurasi 64,3% secara *zero-shot* dan meningkat menjadi 71,2% secara *few-shot*. Temuan ini menjadi landasan teori krusial bagi penelitian ini, karena membuktikan secara empiris bahwa pemberian contoh percakapan nyata dalam *prompt* (*few-shot*) menghasilkan respon yang lebih akurat dan kontekstual dibanding tanpa contoh (*zero-shot*), tanpa memerlukan pembaruan parameter model ataupun *Fine-tuning* tambahan, yang relevan langsung dengan pendekatan pengembangan *chatbot customer service* berbasis LLM dalam penelitian ini[10].

Tabel 1. Perbandingan Studi Sebelumnya

Penulis	Tahun	Metode	Hasil
Meshkin, dkk.	2024	*Zero-shot* vs *Few-shot Prompting* (domain biomedis, Briefings in Bioinformatics Q1).	*Few-shot* secara konsisten mengungguli *Zero-shot* pada domain teknis spesifik; peningkatan F1-score 8–15%; contoh percakapan terbukti sebagai panduan konteks yang efektif untuk LLM di domain teknis.
Sivarajkumar, dkk.	2024	*Zero-shot*, *Few-shot*, dan Heuristic prompts.	Tailoring prompt yang spesifik sangat krusial; GPT-3.5 mengungguli Gemini dan LLaMA-2 dalam konteks klinis.
Reynolds & McDonell	2021	Prompt Programming, *Few-shot* vs *Zero-shot*	Menemukan bahwa desain prompt yang baik (instruksi yang jelas) seringkali lebih efektif daripada memberikan banyak contoh (*few-shot*) yang tidak terstruktur.
Haromain, dkk.	2025	*Zero-shot* (prompt bebas), *Few-shot* (prompt terbimbing)	Analisis prompt *engineering* LLM RAG; *few-shot* unggul di BERTScore/ROUGE untuk info obat.
Brown, dkk.	2020	*Zero-shot*, One-*shot*, dan *Few-shot prompting* pada GPT-3 (175B parameter)	*Few-shot prompting* secara konsisten mengungguli *zero-shot* pada 20+ benchmark NLP tanpa fine-tuning, dengan peningkatan performa yang semakin signifikan seiring skala model contohnya pada TriviaQA, akurasi meningkat dari 64.3% (*zero-shot*) menjadi 71.2% (*few-shot*).

Kelima studi sebelumnya menguji *Zero-shot* dan *Few-shot Prompting* pada domain biomedis, klinis, atau farmasi menggunakan metrik evaluasi otomatis seperti F1-score, BERTScore, dan ROUGE, tanpa melibatkan pengguna nyata dalam proses penilaian. Belum terdapat studi yang secara spesifik membandingkan kedua metode tersebut dalam konteks *chatbot customer service* untuk perusahaan distribusi produk teknis di Indonesia. Penelitian ini mengisi kesenjangan tersebut dengan membandingkan kedua metode pada data kebijakan aktual PT. Samudra Teknik Sejahtera, dievaluasi melalui evaluasi mandiri lima metrik teknis dan evaluasi triangulasi yang melibatkan pengguna nyata.

2.2. Dasar Teori

2.2.1. Artificial Intelligence

Artificial Intelligence (AI) merupakan cabang ilmu komputer yang menekankan penciptaan sistem serta teknologi mampu meniru atau mereplikasi kecerdasan manusia. AI melibatkan beragam metode dan algoritma yang memungkinkan mesin belajar dari data, memproses bahasa alami, mendeteksi pola, serta mengambil keputusan. Di bidang bisnis, AI sangat berperan dalam menyempurnakan pengalaman pelanggan melalui layanan yang lebih personal dan cepat tanggap. AI khususnya meningkatkan kepuasan pelanggan di industri perbankan dan komunikasi jarak jauh, sambil menunjukkan hubungan searah yang positif antara AI dan pengalaman pelanggan. Walaupun AI memberikan keuntungan seperti peningkatan kepuasan pelanggan dan efisiensi operasional, hambatan seperti biaya penerapan, isu privasi data, serta batasan pengolahan bahasa perlu diatasi agar potensinya maksimal[17].

2.2.2. Large Language Model (LLM)

Large Language Model (LLM) adalah model pembelajaran mesin berbasis arsitektur teknologi pemaham konteks kalimat yang dilatih pada dataset teks yang sangat besar untuk memahami, menghasilkan, dan memanipulasi bahasa natural. LLM seperti GPT (Generative Pre-trained Transformer), Claude, dan model serupa lainnya memiliki kemampuan untuk melakukan berbagai tugas

NLP tanpa perlu pelatihan khusus untuk setiap tugas, yang dikenal sebagai *In-context learning*.

Karakteristik utama LLM mencakup skala besar, di mana model memiliki miliaran hingga triliunan parameter yang memungkinkan mereka menangkap pola bahasa yang kompleks. LLM melalui fase *Pre-training* di mana model dilatih pada kumpulan teks yang sangat besar dari internet, buku, dan sumber lainnya untuk mempelajari representasi bahasa yang umum. Kemampuan *few-shot* dan *zero-shot learning* memungkinkan LLM melakukan tugas baru dengan sedikit atau tanpa contoh spesifik, hanya dengan instruksi dalam bentuk *prompt*. Transfer learning memungkinkan pengetahuan yang dipelajari dari *Pre-training* dapat ditransfer ke berbagai tugas spesifik tanpa perlu pelatihan ulang yang menyeluruh dan memakan waktu. LLM juga memiliki contextual understanding yang memungkinkan model memahami konteks dalam percakapan dan dapat mempertahankan koherensi dalam dialog yang panjang[10].

2.2.3. Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) adalah pendekatan yang menggabungkan kemampuan retriever (pencari informasi) dengan generator (LLM) untuk menghasilkan jawaban. RAG dikembangkan untuk mengatasi keterbatasan LLM yang pengetahuannya terbatas pada data pelatihan dan rentan menghasilkan informasi tidak akurat (halusinasi) [18].

Secara teknis, RAG bekerja dengan mengubah dokumen pengetahuan dan pertanyaan pengguna menjadi vektor numerik melalui model embedding, lalu menyimpannya dalam vector store. Saat pengguna mengajukan pertanyaan, sistem mencari potongan dokumen (chunks) yang paling relevan berdasarkan kemiripan vektor menggunakan parameter Top K, kemudian mengirimkannya bersama pertanyaan asli ke LLM sebagai konteks tambahan. Pendekatan ini memastikan jawaban berpijak pada data nyata, bukan hanya memori pelatihan, sehingga mengurangi risiko halusinasi dan meningkatkan akurasi [18].

Dalam penelitian ini, RAG diterapkan sebagai arsitektur utama *chatbot* menggunakan Flowise. Knowledge base PT. Samudra Teknik Sejahtera dipecah

menjadi chunks, diubah menjadi vektor, dan disimpan dalam In-Memory Vector Store. Setiap pertanyaan pelanggan akan mengambil dua potongan dokumen paling relevan (Top K = 2) yang digabungkan dengan prompt (*Zero-Shot* atau *Few-Shot*) sebelum diproses oleh LLM melalui Groq API.

2.2.4. Chatbot

Chatbot adalah aplikasi komputer yang diciptakan untuk meniru obrolan ala manusia. Sebagai solusi canggih di layanan konsumen, *chatbot* ini bisa memperkaya pengalaman berbelanja di platform e-commerce, contohnya SuperAgent yang dibuat spesial untuk website jual-beli online. SuperAgent mengolah data publik secara luas dari situs produk, seperti spesifikasi barang, sesi tanya-jawab konsumen, plus testimoni pembeli. Berbeda dari *chatbot* lama yang terpaku pada interaksi manusia, SuperAgent bijak menangani masalah volume data dan kerahasiaan dengan sumber informasi terbuka dari publik. Sistemnya mencakup beragam komponen: penjawab fakta soal produk, pengurai FAQ pelanggan, penarik kesimpulan opini dari review, serta bagian obrolan ringan. Berbentuk program tambahan browser, SuperAgent memungkinkan akses instan dan pas ke detail produk, sehingga konsumen leluasa menyaring lautan data. Hasil pengujian usability mengonfirmasi manfaatnya: kemudahan meraih info gabungan yang mempercepat alur belanja digital[19].

2.2.5. Prompting

Prompting merupakan instrumen sangat penting untuk mengarahkan *Large Language Model* (LLM) agar menghasilkan *output* yang relevan. Keberhasilan model sangat bergantung pada desain instruksi yang diberikan; riset menunjukkan bahwa variasi pilihan kata dalam sebuah *prompt* dapat menyebabkan perubahan akurasi hingga 17 poin persentase (misalnya, perbedaan antara akurasi 70% dan 53%). Oleh karena itu, pengujian berbagai variasi *prompt* sangat penting untuk mendapatkan performa yang optimal.

Kualitas *output* juga dipengaruhi oleh aspek gaya bahasa, seperti struktur kalimat dan pemilihan kata. Secara umum, instruksi yang lebih sederhana dan jernih cenderung memberikan hasil yang lebih stabil. Perlu dicatat bahwa efektivitas *prompt* tidaklah universal; sebuah perintah yang bekerja sempurna

pada satu model belum tentu efektif pada model lain, sehingga diperlukan penyesuaian (tuning) berbasis konteks spesifik[20].

Metode-Metode *Prompting* yang Umum Digunakan Terdapat berbagai metodologi dalam menyusun perintah AI, yang dikategorikan berdasarkan cara kerja dan tujuan penggunaannya:

a. *Zero-shot Prompting*: Model langsung diberikan tugas atau pertanyaan tanpa adanya contoh referensi sebelumnya.

b. *Few-shot Prompting*: Memberikan beberapa contoh pasangan *input-output* (exemplars) di dalam *prompt* sebagai panduan bagi model sebelum mengerjakan tugas utama.

c. Chain-of-Thought (CoT): Menginstruksikan model untuk memaparkan alur penalaran logisnya secara bertahap sebelum memberikan jawaban final.

d. Tree-of-Thought (ToT): Pengembangan dari CoT di mana model mengeksplorasi beberapa jalur solusi (percabangan) sekaligus, mengevaluasinya, dan memilih jalur terbaik layaknya pohon keputusan.

e. Role-Based *Prompting*: Memberikan identitas atau persona tertentu pada AI untuk mengontrol gaya bahasa, sudut pandang, serta format jawaban.

f. ReAct *Prompting*: Integrasi antara penalaran (*reasoning*) dan tindakan (*acting*). Model akan berpikir terlebih dahulu, melakukan aksi (seperti memanggil API atau mencari data), lalu melanjutkan proses berpikirnya berdasarkan hasil aksi tersebut.

g. Self-Consistency *Prompting*: Strategi di mana model menghasilkan beberapa jalur penalaran yang berbeda secara mandiri, kemudian jawaban akhir ditentukan berdasarkan konsistensi mayoritas dari hasil-hasil tersebut.

Selain daftar di atas, masih banyak teknik *prompting* lainnya yang terus berkembang untuk memenuhi kebutuhan spesifik pengguna di berbagai domain aplikasi[21].

2.2.6. Zero-shot Prompting

Zero-shot Prompting adalah teknik di mana LLM diminta untuk melakukan tugas tanpa diberikan contoh spesifik sebelumnya. Model hanya mengandalkan instruksi umum dan pengetahuan yang telah dipelajarinya selama fase *Pre-training* untuk menghasilkan respon.

Karakteristik *Zero-shot Prompting* mencakup tidak adanya contoh di mana model tidak diberikan demonstrasi atau contoh tentang bagaimana melakukan tugas. *Prompt* berisi instruksi langsung yang eksplisit tentang apa yang harus dilakukan. Model mengandalkan pengetahuan bawaan yang sudah dipelajari selama *Pre-training*. *Zero-shot Prompting* memiliki fleksibilitas tinggi karena dapat digunakan untuk berbagai tugas tanpa perlu mengumpulkan contoh.

Struktur umum *Zero-shot Prompt* terdiri dari instruksi tugas, konteks, dan pertanyaan atau input. Sebagai contoh untuk *customer service*: "Kamu adalah asisten *customer service* PT. Samudra Teknik Sejahtera, distributor alat pertukangan, perkakas, dan sanitary. Jawab pertanyaan pelanggan dengan profesional dan akurat. Pertanyaan: Apakah PT. Samudra Teknik Sejahtera menjual produk OWNER?"

Kelebihan *Zero-shot Prompting* meliputi implementasi cepat karena tidak perlu waktu untuk mengumpulkan dan menyiapkan contoh, skalabilitas yang memungkinkan penerapan mudah untuk berbagai jenis pertanyaan tanpa modifikasi besar, fleksibilitas dalam menangani pertanyaan yang tidak terduga, dan efisiensi token karena menggunakan lebih sedikit token tanpa contoh.

Kekurangan *Zero-shot Prompting* mencakup konsistensi terbatas di mana respon bisa bervariasi untuk pertanyaan serupa, gaya bahasa generic yang cenderung menghasilkan respon yang kurang spesifik untuk brand, akurasi

terbatas untuk tugas yang sangat spesifik, dan halusinasi risk atau risiko lebih tinggi model menghasilkan informasi yang tidak akurat[8].

2.2.7. Few-shot Prompting

Few-shot Prompting adalah teknik di mana LLM diberikan beberapa contoh nyata tentang bagaimana melakukan tugas sebelum diminta untuk memproses *input* baru. Contoh-contoh ini berfungsi sebagai pengondisian awal yang membantu model memahami pola, format, dan gaya respon yang diharapkan.

Karakteristik *Few-shot Prompting* mencakup contoh demonstrasi di mana model diberikan 1-5 contoh (biasanya) tentang *input* dan *output* yang diharapkan. *In-context learning* memungkinkan model belajar dari contoh yang diberikan dalam konteks *prompt* tanpa update parameter. Pattern recognition memungkinkan model mengidentifikasi pola dari contoh dan menerapkannya pada input baru. Konsistensi format membantu menjaga konsistensi format dan gaya respon.

Struktur umum *Few-shot Prompt* terdiri dari instruksi tugas, diikuti dengan beberapa contoh, kemudian pertanyaan atau input baru. Sebagai contoh untuk *customer service*: "Kamu adalah asisten *customer service* PT. Samudra Teknik Sejahtera. Contoh 1: Pertanyaan: Berapa lama PT. Samudra beroperasi? Jawaban: PT. Samudra Teknik Sejahtera telah beroperasi selama lebih dari 20 tahun sebagai distributor terpercaya di Indonesia. Contoh 2: Pertanyaan: Apa saja merek yang dijual? Jawaban: Kami menjual berbagai merek berkualitas, dengan merek unggulan kami adalah OWNER dan YO-ZURI untuk kebutuhan alat pertukangan dan perkakas. Pertanyaan: Apakah melayani pengiriman ke luar kota?"

Kelebihan *Few-shot Prompting* meliputi akurasi lebih tinggi karena contoh membantu model memahami tugas dengan lebih baik, konsistensi gaya yang menjaga konsistensi tone dan format respon sesuai contoh, kontrol lebih baik yang memberikan kontrol lebih besar atas jenis respon yang dihasilkan, dan

adaptasi domain yang mudah disesuaikan dengan terminologi dan gaya spesifik perusahaan.

Kekurangan *Few-shot Prompting* mencakup memerlukan contoh berkualitas di mana kualitas output sangat bergantung pada kualitas contoh, konsumsi token lebih banyak karena menyertakan contoh, waktu setup lebih lama yang memerlukan waktu untuk menyusun contoh yang representatif, dan maintenance kompleks yang perlu update contoh saat ada perubahan informasi perusahaan[15].

2.2.8. Customer service

Customer service merupakan bagian yang sangat penting dalam berbagai industri, yang mencakup semua tindakan yang diambil untuk membangun dan mempertahankan hubungan dengan pelanggan, dari penyediaan informasi produk hingga penyelesaian masalah pelanggan dengan cepat dan tepat. Dalam era digital, *chatbot* berbasis kecerdasan buatan telah menjadi komponen integral dari strategi *customer service* modern, memungkinkan perusahaan untuk mengotomatisasi tugas-tugas rutin, mengurangi biaya operasional, dan meningkatkan efisiensi layanan. Bocklisch, dkk. dalam penelitiannya tentang pengaruh pemilihan *Large Language Model* (LLM) dan *prompt engineering* terhadap efektivitas *chatbot customer service* menemukan bahwa desain *prompt* yang spesifik sangat mempengaruhi kualitas respon *chatbot*, di mana penggunaan format *prompt* terstruktur seperti JSON dapat meningkatkan akurasi hingga 21,62 poin dibandingkan dengan *prompt* format bebas [22].

Penelitian tersebut menggunakan dataset yang terdiri dari 400 frasa uji berdasarkan percakapan *customer service* aktual dengan pelanggan operator telepon seluler, dan menemukan bahwa model yang lebih kecil seperti Gemini-1.5-Flash-8B dapat mencapai hasil yang sebanding dengan model yang lebih besar jika dikombinasikan dengan teknik *prompt engineering* yang tepat, menawarkan solusi yang hemat biaya untuk implementasi *chatbot customer service*. *Customer service* yang efektif melalui *chatbot* berbasis LLM memerlukan komponen kunci seperti Natural Language Understanding untuk memahami maksud serta tujuan pelanggan, pengelolaan dialog untuk mengelola alur

percakapan, knowledge base yang berisi informasi akurat tentang produk dan layanan, serta *prompt engineering* yang optimal untuk memandu LLM menghasilkan respon yang konsisten dengan identitas merek perusahaan.

Penelitian ini menekankan pentingnya penggunaan format *prompt* terstruktur (YAML, JSON) dan perintah yang presisi untuk meningkatkan performa *chatbot*, serta pentingnya evaluasi berdasarkan data nyata dengan dataset percakapan *customer service* yang sesuai dengan keadaan asli. Dalam konteks pengembangan *chatbot* berbasis LLM, pemilihan metode *prompting* yang tepat baik *zero-shot* maupun *few-shot* dan desain *prompt* yang unggul menjadi krusial untuk memastikan *chatbot* dapat memberikan respon yang akurat, konsisten, dan relevan, sehingga memenuhi atau melampaui ekspektasi pelanggan dalam hal kualitas layanan.