

BAB II

TINJAUAN PUSTAKA

2.1. Studi Sebelumnya

Kekäläinen melakukan penelitian mengenai implementasi teknik *prompting* pada AI *chatbot* layanan pelanggan di sektor pariwisata, dengan fokus membandingkan *zero-shot*, *few-shot prompting*, dan *chain-of-thought prompting*. Penelitian ini bertujuan untuk menguji dampak teknik *prompting* yang berbeda terhadap kualitas layanan *chatbot*. Penelitian ini dilakukan dengan mengukur kemampuan model dalam menjawab pertanyaan pelanggan nyata dari industri pariwisata. Hasil penelitian menunjukkan bahwa *few-shot prompting* dapat digunakan untuk pertanyaan sederhana karena model dapat mengenali pola dari contoh yang diberikan, sementara *chain-of-thought prompting* lebih unggul dalam menangani pertanyaan kompleks yang memerlukan penalaran bertahap. Teknik ini memungkinkan *chatbot* untuk memecah masalah menjadi langkah-langkah kecil sehingga menghasilkan respons yang lebih akurat dan komprehensif. Lebih lanjut, penelitian ini menekankan pentingnya pemilihan teknik *prompting* yang tepat sesuai dengan jenis pertanyaan dan kompleksitas masalah yang dihadapi oleh pelanggan. Dengan demikian, implementasi yang tepat dapat meningkatkan layanan pelanggan berbasis AI[7].

Logan dan rekannya melakukan penelitian mengenai *few-shot learning* dalam pemrosesan bahasa alami, khususnya dalam mengurangi ketergantungan terhadap *prompt engineering* yang kompleks dan bergantung pada intuisi manusia. Mereka menemukan bahwa dengan menggunakan *fine-tuning* pada model bahasa, kebutuhan akan perancangan *prompt* dapat diminimalkan, bahkan null *prompts* yang tidak mengandung pola atau contoh spesifik dapat mencapai akurasi setara dengan *prompt* manual. Selain itu, penelitian ini menunjukkan bahwa metode BitFit, yang hanya memperbarui bias *terms*, dapat menghasilkan kinerja yang kompetitif dengan konsumsi memori yang jauh lebih rendah dibandingkan *fine-tuning* konvensional. Hasil akhir penelitian ini menegaskan bahwa *prompt-based fine-tuning* adalah pendekatan yang lebih akurat, lebih tahan terhadap variasi *prompt*, serta dapat dibuat hampir sebaik

model beku (frozen LMs), sehingga menjadi solusi dalam penerapan *chain-of-thought* learning[8].

Sultan dan rekan-rekannya melakukan penelitian mengenai pengembangan teknik *structured chain-of-thought* (SCoT) *prompting* untuk menghasilkan percakapan tanya jawab (*question-answering*) berbasis konten menggunakan *large language models*. Tujuan penelitian ini adalah untuk meningkatkan kemampuan model dalam menangani tugas-tugas kompleks dengan memecah proses menjadi langkah-langkah terstruktur menggunakan pendekatan berbasis state machine. Penelitian ini menunjukkan bahwa SCoT *prompting* meningkatkan akurasi respons *chatbot* dalam skenario layanan pelanggan yang memerlukan penalaran logis bertahap. Dengan pendekatan ini, *chatbot* dapat memberikan jawaban yang lebih relevan dan sesuai dengan kebutuhan pengguna dibandingkan metode *prompting* lainnya. Selain itu, SCoT juga memungkinkan model untuk mengelola percakapan yang lebih panjang dan kompleks, sehingga meningkatkan pengalaman pengguna secara keseluruhan. Penelitian ini memberikan kontribusi signifikan dalam pengembangan *chatbot* yang lebih cerdas dan adaptif untuk berbagai aplikasi[9].

Kojima dan rekan-rekannya melakukan penelitian tentang kemampuan model bahasa besar untuk melakukan penalaran tanpa pelatihan tambahan (*zero-shot reasoning*), serta membandingkannya dengan pendekatan *chain-of-thought prompting* dan *chain-of-thought prompting*. Penelitian ini bertujuan untuk mengevaluasi berbagai teknik *prompting* dalam menyelesaikan tugas-tugas kompleks seperti layanan pelanggan otomatis dan pemecahan masalah logika. Hasil penelitian menunjukkan bahwa *chain-of-thought prompting* secara signifikan meningkatkan akurasi model pada tugas-tugas yang membutuhkan penalaran bertahap dibandingkan teknik lainnya. Teknik ini memungkinkan model untuk memecah masalah menjadi langkah-langkah kecil sehingga menghasilkan jawaban yang lebih logis dan terstruktur. Implikasi dari penelitian ini adalah bahwa *chain-of-thought prompting* dapat menjadi solusi untuk meningkatkan kinerja *chatbot* layanan pelanggan yang membutuhkan kemampuan

penalaran yang kompleks. Teknik ini memungkinkan *chatbot* untuk memberikan jawaban yang lebih akurat dan memuaskan bagi pengguna[10].

Zhou dan timnya melakukan penelitian tentang pengembangan teknik *prompting* baru bernama *least-to-most prompting*, yang dikombinasikan dengan pendekatan *chain-of-thought prompting* untuk menangani tugas-tugas kompleks menggunakan *large language models*. Penelitian ini bertujuan untuk meningkatkan kemampuan *reasoning* model AI dengan memulai dari langkah-langkah sederhana menuju langkah-langkah yang lebih kompleks secara bertahap. Hasil penelitian menunjukkan bahwa kombinasi kedua teknik ini mampu meningkatkan akurasi respons model sebesar 25% pada tugas-tugas yang membutuhkan analisis mendalam dibandingkan dengan metode *prompting* tradisional seperti *chain-of-thought*. Dengan demikian, penelitian ini memberikan wawasan baru tentang cara meningkatkan kinerja model AI dalam memecahkan masalah yang kompleks. Teknik *least-to-most prompting* dan *chain-of-thought prompting* dapat diintegrasikan dalam pengembangan *chatbot* layanan pelanggan untuk memberikan respons yang lebih akurat dan memuaskan[11].

Tabel 1 Tabel Perbandingan Studi Sebelumnya

Penulis	Tahun	Metode	Hasil
Kekäläinen, dkk	2024	Zero-shot, <i>Chain-of-thought</i> , <i>Chain-of-thought prompting</i>	<i>Few-Shot</i> unggul untuk pertanyaan sederhana, <i>Chain-of-thought</i> unggul dalam pertanyaan kompleks.
Logan, dkk	2021	<i>Chain-of-thought learning</i> , null <i>prompts</i>	<i>Chain-of-thought learning</i> dapat dioptimalkan dengan <i>prompt-based fine-tuning</i> , yang mengurangi ketergantungan pada <i>prompt engineering</i> . Null <i>prompts</i> terbukti memiliki akurasi setara dengan <i>prompt</i> manual, sehingga menyederhanakan desain <i>prompt</i> tanpa menurunkan performa model.
Sultan, dkk	2024	Structured <i>Chain-of-thought Prompting</i> (SCoT)	SCoT meningkatkan akurasi <i>chatbot</i> dalam layanan pelanggan yang membutuhkan penalaran logis bertahap
Kojima, dkk	2023	Zero-shot, <i>Chain-of-thought</i> , <i>Chain-of-thought prompting</i>	<i>Chain-of-thought</i> meningkatkan akurasi model dalam tugas yang memerlukan penalaran bertahap
Zhou, dkk	2023	Least-to-Most <i>Prompting</i> & <i>Chain-of-thought Prompting</i>	Kombinasi teknik meningkatkan akurasi respons model sebesar 25% dalam tugas yang kompleks

2.2. Dasar Teori

2.2.1. Artificial Intelligence

Artificial intelligence (AI) adalah cabang ilmu komputer yang berfokus pada pengembangan sistem dan teknologi yang dapat meniru atau mensimulasikan kecerdasan manusia. AI mencakup berbagai teknik dan algoritma yang memungkinkan mesin untuk belajar dari data, memahami bahasa alami, mengenali pola, dan membuat keputusan. Dalam konteks bisnis, AI berperan penting dalam meningkatkan pengalaman pelanggan dengan menyediakan layanan yang lebih personal dan responsif. AI dapat meningkatkan pengalaman pelanggan, terutama di sektor perbankan dan telekomunikasi, serta mengeksplorasi hubungan positif antara AI dan pengalaman pelanggan. Meskipun AI menawarkan banyak manfaat, seperti peningkatan kepuasan pelanggan dan operasional, tantangan seperti biaya implementasi, kekhawatiran privasi data, dan keterbatasan dalam pemrosesan bahasa masih perlu diatasi untuk memaksimalkan potensi teknologi ini[12].

2.2.2. Customer Service dan Sales Otomatis

Penggunaan chatbot dalam konteks otomasi layanan pelanggan (*Customer service Automation*) dan penjualan (*Sales Automation*) memiliki potensi besar sebagai solusi untuk meningkatkan kinerja operasional perusahaan. Berdasarkan temuan dalam studi berjudul *Chatbots and Customer Service: Enhancing Marketing Management with AI Technologies*, chatbot AI dapat mengotomatisasi tugas-tugas rutin dalam layanan pelanggan, seperti menjawab pertanyaan umum dan mengarahkan pelanggan, sehingga memungkinkan agen manusia berkonsentrasi pada kasus yang lebih kompleks. Studi tersebut juga menekankan pentingnya desain yang berpusat pada pengguna (*user-centric design*) untuk memastikan interaksi chatbot yang intuitif, responsif, dan memuaskan, yang berkontribusi pada peningkatan loyalitas dan kepercayaan pelanggan[26]. Dalam kerangka *sales automation*, otomasi ini memungkinkan proses sapaan awal, pemberian rekomendasi produk, dan perekapan data pembelian dapat berjalan otomatis, mempercepat waktu respons serta meningkatkan peluang konversi penjualan. Secara teori, integrasi CS dan Sales

Automation melalui chatbot AI selaras dengan tujuan strategis perusahaan untuk meningkatkan kualitas interaksi pelanggan sekaligus memaksimalkan potensi penjualan otomatis.

2.2.3. Chatbot

Chatbot adalah program komputer yang dirancang untuk mensimulasikan percakapan manusia. *Chatbot* dapat digunakan sebagai solusi inovatif dalam layanan pelanggan untuk meningkatkan pengalaman belanja online, seperti yang ditunjukkan oleh *SuperAgent*, sebuah *chatbot* khusus untuk situs web *e-commerce*. *SuperAgent* memanfaatkan data berskala besar yang tersedia secara publik dari halaman produk, termasuk deskripsi produk, pertanyaan dan jawaban pelanggan, serta ulasan. Berbeda dengan *chatbot* tradisional yang bergantung pada dialog manusia, *SuperAgent* mengatasi tantangan skala data dan privasi dengan memanfaatkan informasi yang dikumpulkan dari masyarakat. *Chatbot* ini terdiri dari beberapa mesin, termasuk mesin penjawab pertanyaan fakta untuk informasi produk, mesin pencarian FAQ untuk pertanyaan pelanggan, mesin penambangan opini untuk ulasan, dan mesin *chit-chat* untuk percakapan santai. Diimplementasikan sebagai ekstensi browser, *SuperAgent* memberikan akses cepat dan relevan ke informasi produk, sehingga memudahkan pelanggan dalam menavigasi volume besar data terkait produk. Analisis kegunaan menunjukkan bahwa pelanggan mendapatkan manfaat dari akses mudah ke informasi yang teragregasi, meningkatkan proses belanja online[13].

2.2.4. Large Language Models

Large Language Models atau dapat disingkat sebagai LLM, yang merupakan model bahasa canggih yang dilatih menggunakan data dalam jumlah besar untuk memahami dan menghasilkan teks dengan cara yang menyerupai manusia. Sejak kemunculannya, LLM telah merevolusi teknologi *chatbot*, dengan contoh-contoh terkenal seperti ChatGPT, Google's BARD (sekarang Gemini), dan Microsoft's Bing Chat. Model-model ini memiliki kemampuan unik, seperti akses informasi *real-time*, gaya respons yang dapat disesuaikan, dan kemampuan penalaran yang lebih maju. LLM digunakan dalam berbagai sektor, termasuk pendidikan, di mana mereka

meningkatkan pengalaman belajar dan mendukung penulisan akademis, penelitian di mana mereka mempercepat tinjauan literatur dan analisis data, serta kesehatan di mana mereka memberikan dukungan untuk pertanyaan medis dan perawatan pasien. Meskipun LLM menawarkan banyak manfaat, tantangan seperti pembaruan pengetahuan, bias dalam data pelatihan, dan masalah etika tetap menjadi perhatian penting yang perlu diatasi untuk memastikan penggunaan yang bertanggung jawab dari teknologi ini[14].

2.2.5. Prompting

Prompting adalah teknik yang digunakan untuk memandu LLM dalam menghasilkan respons yang diinginkan. Kinerja model sangat sensitif terhadap pilihan *prompt*, dengan variasi kinerja yang dapat mencapai 17 poin persentase, yang berarti bahwa perbedaan dalam akurasi antara *prompt* yang berbeda bisa sangat signifikan, misalnya, satu *prompt* dapat menghasilkan akurasi 70% sementara yang lain hanya 53%. Ini menunjukkan pentingnya menguji berbagai *prompt* untuk mendapatkan estimasi kinerja yang lebih akurat. Selain itu, properti linguistik dari *prompt*, seperti struktur kalimat dan frekuensi kata, juga berpengaruh terhadap hasil yang diperoleh, di mana *prompt* yang lebih sederhana dan jelas cenderung memberikan kinerja yang lebih baik. Penelitian menunjukkan bahwa *prompt* tidak bersifat universal, sebuah *prompt* yang optimal untuk satu model mungkin tidak cocok untuk model lain, sehingga penyesuaian berdasarkan konteks spesifik sangat diperlukan[15].

Terdapat banyak sekali jenis metode *prompting* yang dapat digunakan untuk memberikan perintah kepada AI. Berikut adalah beberapa jenis metode prompting AI

- a. *Zero-Shot Prompting*, metode di mana model diminta menyelesaikan tugas tanpa diberi contoh sebelumnya. Hanya ada instruksi atau pertanyaan, lalu model menjawab langsung.
- b. *Few-shot Prompting*, metode dengan model yang diberi beberapa contoh (*exemplars*) dalam prompt untuk menunjukkan cara menyelesaikan tugas, lalu diminta mengerjakan kasus baru.

- c. *Chain-of-thought Prompting*, metode yang membuat model menuliskan proses berpikirnya langkah demi langkah sebelum memberikan jawaban akhir.
- d. *Tree-of-thought Prompting*, perluasan dari CoT, tapi model menghasilkan beberapa jalur kemungkinan langkah (branching) seperti pohon keputusan. Setiap cabang dievaluasi, dan hanya jalur terbaik yang dilanjutkan.
- e. *Role-based prompting*, memberi peran atau persona tertentu pada model yang dapat membantu mengontrol gaya bahasa, perspektif, dan format output.
- f. *ReAct prompting*, menggabungkan *reasoning* (penalaran) dan *acting* (aksi). Model berpikir secara bertahap, lalu mengambil tindakan (misalnya memanggil API, mencari data) sebelum melanjutkan ke langkah berikutnya.
- g. *Self-consistency prompting*, meminta model membuat beberapa jalur penalaran berbeda lalu memilih jawaban akhir berdasarkan mayoritas hasil yang konsisten.

Dan masih banyak lagi jenis *prompt* yang setiap metodenya dikelompokkan berdasarkan cara penggunaannya sesuai dengan kebutuhan dari masing-masing pengguna AI[16].

2.2.6. Few-Shot Prompting

Few-shot prompting, atau *FW-prompting*, merupakan pendekatan dalam pembelajaran mesin yang memungkinkan model bahasa besar (LLM) untuk belajar dari sejumlah kecil contoh (shots) yang diberikan dalam *prompt*. Dalam konteks ini, terdapat dua paradigma utama yaitu *Explain-then-Predict* (E-P) dan *Predict-then-Explain* (P-E). Pada E-P, penjelasan disediakan sebelum prediksi, sedangkan pada P-E, penjelasan dihasilkan setelah prediksi dilakukan. Penelitian menunjukkan bahwa penggunaan penjelasan dapat meningkatkan akurasi model dalam berbagai tugas, dengan E-P sering kali menunjukkan performa yang lebih baik dibandingkan P-E, terutama ketika dikombinasikan dengan teknik kalibrasi. Dengan memanfaatkan penjelasan yang jelas dan terstruktur, *FW-prompting* dapat membantu model dalam

memahami dan memproses informasi, meskipun hanya dengan sedikit data pelatihan[17].

2.2.7. Chain-of-Thought Prompting

Chain-of-thought (CoT) prompting adalah teknik yang digunakan untuk mendorong model bahasa besar (LLM) dalam menghasilkan langkah-langkah penalaran yang mendahului jawaban akhir. Teknik ini terbagi menjadi dua paradigma utama: Auto-CoT, yang menggunakan *prompt* sederhana seperti "Mari berpikir langkah demi langkah" tanpa contoh sebelumnya, dan Manual-CoT, yang memerlukan demonstrasi yang ditulis secara manual untuk setiap pertanyaan. CoT *prompting* memungkinkan LLM untuk memecah masalah kompleks menjadi langkah-langkah yang lebih sederhana, sehingga meningkatkan kemampuan penalaran mereka. Namun, Manual-CoT memerlukan upaya manual yang signifikan dalam mendesain demonstrasi, yang dapat diatasi dengan pendekatan Auto-CoT, yang secara otomatis membangun demonstrasi untuk meningkatkan kinerja penalaran[18].