Analisis Komparatif Performa *Large Language Model* dalam Perancangan *User Interface* Berbasis *Framework* CO-STAR

Jovan Nathaniel Jie[1)], Kristopher Owen[2)], Michael Jovanco[3)]

[1)2)3)4)]Sistem Informasi, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang
Jl. Rajawali No. 14, 9 Ilir, Kec. Ilir Tim. II, Sumatera Selatan, Palembang

[1)]jovannathanieljie_2327240013@mhs.mdp.ac.id
[2)]kristopherowen_2327240060@mhs.mdp.ac.id
[3)]michaeljovanco_2327240023@mhs.mdp.ac.id
[4)]dqpratama@mdp.ac.id

Abstrak

Penelitian ini mengeksplorasi strategi perancangan antarmuka pengguna (*user interface*) yang intuitif bagi masyarakat awam dengan tingkat literasi digital terbatas melalui pemanfaatan kecerdasan buatan. Fokus utama studi ini adalah menganalisis bagaimana kualitas instruksi dalam prompt engineering memengaruhi efektivitas *Large Language Model* (LLM) dalam menghasilkan solusi desain. Peneliti mengusulkan penggunaan *framework* CO-STAR sebagai struktur instruksi formal untuk meningkatkan kejelasan dan relevansi *output* AI. Melalui eksperimen komparatif terhadap enam model LLM, hasil desain dievaluasi oleh 129 responden menggunakan skala Likert dan analisis statistik *Mean Rank*. Temuan menunjukkan bahwa kerangka kerja CO-STAR secara signifikan membantu AI dalam menyederhanakan elemen visual yang kompleks menjadi tata letak yang mudah dipahami oleh pengguna pemula. Studi ini menyimpulkan bahwa rekayasa prompt yang presisi mampu mengatasi limitasi sistem tokenisasi AI, sekaligus memberikan kontribusi praktis dalam menciptakan produk digital yang mendukung inklusivitas digital bagi seluruh lapisan masyarakat.

Kata kunci: *Large Language Model*, *Prompt Engineering*, CO-STAR, *User Interface*, Masyarakat Awam

Abstract

This research investigates strategies for designing intuitive user interfaces for individuals with limited digital literacy by leveraging artificial intelligence technology. The primary focus of this study is to analyze how the quality of instructions in prompt engineering affects the effectiveness of Large Language Models (LLMs) in generating accessible design solutions. The researchers propose the implementation of the CO-STAR framework as a formal instructional structure to enhance the clarity and relevance of AI output. Through a comparative experiment involving six LLM models, the design outcomes were evaluated by 129 respondents using a Likert scale and Mean Rank statistical analysis. The findings indicate that the CO-STAR framework significantly assists AI in simplifying complex visual elements into layouts that are easily understood by novice users. This study concludes that precise prompt engineering can overcome the limitations of AI tokenization systems while providing a practical contribution to creating digital products that promote digital inclusivity for all levels of society.

Keywords: *Large Language Model, Prompt Engineering, CO-STAR, User Interface, Digital Inclusivity*

1. PENDAHULUAN

1.1. LATAR BELAKANG

Di Indonesia, tantangan inklusivitas digital masih menjadi isu krusial, terutama bagi masyarakat awam yang memiliki tingkat literasi teknologi rendah. Aplikasi layanan publik,

DOI:https://doi.org/10.47111/jointecoms.v6i2 *Received: 01-06-2026*
Accepted: 10-06-2026

seperti "Layanan Mandiri Desa", sering kali dirancang dengan antarmuka pengguna yang terlalu kompleks dan penuh dengan jargon teknis yang mengintimidasi pengguna pemula. Permasalahan nyata yang dihadapi adalah sulitnya masyarakat awam dalam melakukan navigasi mandiri tanpa bantuan pihak lain, yang pada akhirnya menghambat efisiensi pelayanan birokrasi di tingkat desa.

Di sisi lain, teknologi kecerdasan buatan, khususnya *Large Language Model* (LLM), tengah berkembang pesat melalui arsitektur *Transformers* yang memungkinkan mesin memahami konteks bahasa alami manusia secara mendalam. Model-model AI mutakhir seperti GPT, Gemini, Qwen dan Claude kini memiliki kemampuan untuk melakukan tugas-tugas kompleks, mulai dari generasi teks hingga perancangan konsep visual, asalkan diberikan instruksi yang jelas dan terstruktur. Kemunculan disiplin ilmu *Prompt Engineering* menjadi kunci utama dalam mengoptimalkan interaksi antara manusia dan AI agar menghasilkan *output* yang relevan dan akurat.

Penelitian ini sangat dibutuhkan untuk mengetahui sejauh mana peran dan dampak AI dalam memecahkan hambatan desain bagi masyarakat awam melalui pendekatan rekayasa *prompt* yang sistematis. Dengan mengaitkan kebutuhan desain inklusif dan potensi LLM, penelitian ini menggunakan *framework* CO-STAR sebagai "cetak biru" instruksi untuk menjamin kejelasan peran dan tujuan bagi AI. Perbandingan antara berbagai model AI diperlukan untuk mengevaluasi efektivitas masing-masing sistem dalam menangani aspek teknis seperti sistem tokenisasi yang sering kali memengaruhi akurasi teks pada elemen visual. Melalui analisis komparatif ini, diharapkan dapat ditemukan model AI dan strategi *prompting* terbaik yang mampu menghasilkan antarmuka yang paling mudah dipahami oleh masyarakat luas.

1.2. Rumusan Masalah

Berdasarkan latar belakang tersebut, rumusan masalah yang didapatkan dalam penelitian ini adalah sebagai berikut:

1. Mengapa penggunaan *framework* CO-STAR dalam rekayasa *prompt* sangat berpengaruh terhadap kualitas dan relevansi desain antarmuka yang dihasilkan oleh LLM untuk masyarakat awam?
2. Bagaimana perbandingan performa antara model GPT 5.3, Gemini 3.1 Pro, Gemini 3.1 Fast, Claude Haiku 4.5, Claude Sonnet 4.6 dan Qwen 3.6 Plus dalam menghasilkan desain *user interface* yang akurat dan estetis bagi pengguna dengan literasi digital rendah?

1.3. Tujuan Penelitian

1. Menganalisis efektivitas penerapan *framework* CO-STAR dalam meningkatkan performa model AI untuk merancang solusi desain yang inklusif.
2. Membandingkan hasil keluaran dari berbagai model LLM guna menentukan teknologi mana yang paling andal dalam menciptakan antarmuka layanan publik yang ramah bagi masyarakat awam.

1.4. Manfaat Penelitian

1. Bagi akademisi: Memberikan kontribusi pada pengembangan metodologi rekayasa *prompt* dalam bidang Sistem Informasi, khususnya pada topik teknologi berkembang.
2. Bagi praktisi/pengembang: Menjadi referensi dalam memanfaatkan AI untuk mempercepat proses desain antarmuka yang berorientasi pada kemudahan akses bagi seluruh lapisan masyarakat.

2. TINJAUAN PUSTAKA

2.1. Teori *Prompt Engineering* dan *Framework* CO-STAR

Keberhasilan *Large Language Model* (LLM) dalam menghasilkan desain yang relevan bergantung pada kualitas instruksi. Di dalam [5], dijelaskan bahwa katalog pola prompt membantu memitigasi ketidakpastian *output* AI. *Framework* terstruktur seperti CO-STAR berfungsi memberikan batasan kontekstual yang ketat sehingga model tidak menghasilkan *output*

di luar koridor kebutuhan pengguna [6]. Selain itu, teknik *few-shot prompting* terbukti mampu meningkatkan stabilitas AI dalam mengikuti instruksi yang kompleks dan spesifik [10].

2.2. Kemampuan LLM dalam Perancangan UI dan Tata Letak

LLM modern kini memiliki kemampuan penalaran spasial yang memungkinkan perencanaan visual antarmuka. Berdasarkan penelitian dalam [12], model bahasa besar dapat diarahkan untuk menghasilkan tata letak (*layout*) secara komposisional yang logis. Namun, evaluasi dalam [1] menunjukkan adanya variasi performa antarmodel dalam memahami hierarki visual. Selain itu, integrasi AI dalam alur kerja desain menuntut pemahaman mendalam tentang bagaimana model memproses instruksi teknis untuk menghindari kesalahan representasi [8].

2.3. Desain Antarmuka bagi Pengguna Literasi Digital Rendah

Masyarakat awam dengan literasi teknologi rendah memerlukan pendekatan desain yang meminimalkan beban kognitif. Di dalam [3], ditekan bahwa isyarat non-tekstual (seperti ikon besar) jauh lebih efektif daripada teks instruksional. Penggunaan bahasa sehari-hari yang ramah dan bantuan berbasis suara juga terbukti meningkatkan tingkat keberhasilan tugas bagi pengguna pemula [4]. Prinsip desain bagi kelompok ini harus mengutamakan navigasi yang dangkal dan elemen visual yang sangat kontras untuk memudahkan keterbacaan [9].

2.4. Tantangan Tokenisasi dan Akurasi Tekstual AI

Salah satu hambatan teknis utama dalam menggunakan LLM untuk desain visual adalah sistem tokenisasi. Di dalam [2], dijelaskan bahwa AI sering kali mengalami kendala saat harus merender teks spesifik di dalam elemen grafis. Hal ini menyebabkan diperlukannya rekayasa prompt yang sangat detail untuk menjamin ejaan pada tombol atau label UI tetap akurat. Penelitian dalam [11] menyoroti bahwa ketergantungan model pada probabilitas kata terkadang mengabaikan ketepatan visual yang diminta oleh pengguna.

2.5. Inklusivitas Digital dalam Layanan Publik Desa

Transformasi digital di tingkat desa harus mampu menjembatani jurang literasi melalui antarmuka yang empatik. Berdasarkan [7], digitalisasi layanan publik tidak hanya soal ketersediaan teknologi, tetapi soal aksesibilitas universal bagi seluruh lapisan masyarakat. Penggunaan AI untuk menghasilkan prototipe UI yang inklusif merupakan langkah strategis dalam menciptakan sistem layanan mandiri desa yang tidak lagi mengintimidasi masyarakat awam.

3. METODE PENELITIAN

Penelitian ini menerapkan metode kuantitatif komparatif dengan pendekatan eksperimen terkontrol untuk mengevaluasi performa enam model *Large Language Model* (LLM), yaitu GPT 5.3, Gemini 3.1 Pro, Gemini 3.1 Fast, Claude Haiku 4.5, Claude Sonnet 4.6 dan Qwen 3.6 Plus. Langkah awal penelitian dilakukan dengan tahap rekayasa instruksi (*prompt engineering*) menggunakan *framework* CO-STAR untuk merancang antarmuka aplikasi "Layanan Mandiri Desa". *Output* visual yang dihasilkan dari masing-masing model kemudian dikumpulkan dan disajikan kepada responden dalam format kuesioner terstruktur untuk dinilai tingkat kelayakannya berdasarkan parameter desain yang telah ditetapkan.

Penentuan jumlah sampel dalam penelitian ini dilakukan menggunakan Rumus Lemeshow, mengingat total populasi pengguna layanan publik bersifat tidak terbatas atau jumlahnya tidak diketahui secara pasti. Perhitungan dilakukan dengan tingkat kepercayaan 95% (Z = 1,96), estimasi proporsi (P = 0,5), serta tingkat kesalahan atau *margin of error* sebesar 10% (d = 0,10).

$$n = \frac{Z^2 .P(1-P)}{d^2} = \frac{(1{,}96)^2 .(0{,}5)(1-0{,}5)}{(0{,}10)^2} \approx 96{,}04 \quad (1)$$

Berdasarkan perhitungan Persamaan (1), diperoleh hasil sebesar 96,04. Simbol-simbol yang digunakan dalam persamaan tersebut didefinisikan sebagai berikut.
n = jumlah sampel minimal yang diperlukan.
Z = skor statistik berdasarkan tingkat kepercayaan (1,96 untuk tingkat kepercayaan 95%).
P = estimasi proporsi populasi atau prevalensi fenomena yang diteliti.
d = tingkat kesalahan atau presisi yang diinginkan (*margin of error*).

Maka, peneliti menetapkan jumlah sampel minimal yang dibutuhkan sebanyak 96 responden. Namun, jumlah tersebut akan dibulatkan menjadi 100 untuk memperkuat validitas data yang diambil dari masyarakat umum sebagai calon pengguna potensial.

Proses validasi data dilakukan melalui uji reliabilitas instrumen kuesioner guna memastikan konsistensi butir penilaian pada *Google Form*. Pengukuran dilakukan menggunakan skala Likert 1-5 yang mencakup tiga indikator utama: kesesuaian *prompt*, kualitas estetika, dan penilaian keseluruhan tampilan. Penilaian ini bertujuan untuk memperoleh data mentah mengenai model AI mana yang paling akurat dalam menerjemahkan kebutuhan visual tanpa mengalami kesalahan ejaan atau *typo* yang dapat membingungkan masyarakat awam.

Analisis data akhir dilakukan dengan mengolah skor dari setiap model AI pada setiap indikator penilaian. Untuk membuktikan adanya perbedaan performa yang signifikan antara ketiga model tersebut, peneliti menggunakan uji statistik *mean rank* Kruskal-Wallis. Hasil pengolahan data ini akan menjadi dasar penarikan kesimpulan mengenai model LLM terbaik yang mampu menghasilkan antarmuka aplikasi publik yang ramah pengguna, inklusif, dan bebas dari kompleksitas teknis.

4. PEMBAHASAN

4.1. Validasi Sampel dan Reliabilitas

Berdasarkan perhitungan menggunakan Rumus Lemeshow, target minimal sampel yang ditetapkan adalah 96 responden. Dalam pelaksanaan penelitian, kuesioner berhasil mengumpulkan data dari 129 responden (melebihi target minimal 100). Hal ini memperkuat validitas hasil penelitian dan menurunkan *margin of error* sehingga data yang diperoleh representatif untuk merepresentasikan penilaian masyarakat umum terhadap desain antarmuka layanan publik.

4.2. Implementasi *Prompt Engineering* (CO-STAR *Framework*)

Tahap eksperimen dimulai dengan penyusunan instruksi terstruktur menggunakan *framework* CO-STAR untuk memitigasi variasi *output* AI. *Prompt* dirancang khusus untuk menguji kemampuan LLM dalam menciptakan desain inklusif bagi masyarakat dengan literasi digital rendah. Adapun *prompt* yang digunakan adalah sebagai berikut:

> ***CONTEXT* (C):** Anda adalah seorang *Senior* UI/UX *Designer* yang ahli dalam menciptakan aplikasi untuk pengguna pemula dengan literasi teknologi rendah. Saya sedang melakukan penelitian jurnal untuk membandingkan bagaimana berbagai LLM memahami kebutuhan masyarakat awam dalam menggunakan aplikasi layanan publik bernama "Layanan Mandiri Desa".
>
> ***OBJECTIVE* (O):** Tugas utama Anda adalah menghasilkan 3 gambar visual terpisah (*mockup high-fidelity*) yang merepresentasikan antarmuka aplikasi *mobile*. Anda harus menciptakan representasi grafis, bukan deskripsi teks. Halaman yang harus dihasilkan adalah: (1) Halaman Masuk/*Login*, (2) Halaman Pengajuan Surat, dan (3) Halaman Status.
>
> ***STYLE* (S):** Gunakan gaya "*Ultra-Simple & Visual-Heavy*". Visual: Gunakan ikon nyata dan besar. Tipografi: Sans-Serif bersih, *Headline* 30 pt, *Body* 18 pt. Warna:

Kontras tinggi, latar putih, tombol biru terang (#007BFF), bantuan merah terang (#FF0000). Bahasa: Hindari istilah teknis, gunakan bahasa sehari-hari yang ramah.

***TONE* (T):** Sangat membimbing, sabar, dan jelas. Hindari gradien atau bayangan rumit.

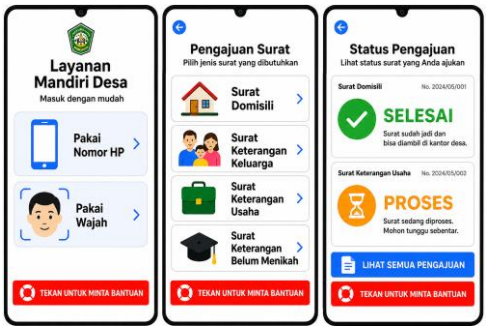
***AUDIENCE* (A):** Masyarakat pedesaan atau warga awam yang jarang menggunakan aplikasi *smartphone.*

***RESPONSE FORMAT* (R):** Hasilkan 3 gambar *mockup* visual secara langsung. WAJIB: Sertakan tombol dengan teks eksak: "**TEKAN UNTUK MINTA BANTUAN**" untuk menguji sistem tokenisasi.

BATASAN KERAS: Jangan berikan deskripsi teks panjang, kode program, atau penjelasan alasan desain. Hanya hasilkan gambar visual.

4.3. Analisis Visual Hasil Desain UI per Model LLM

Berdasarkan *prompt* di atas, didapatkan hasil rancangan antarmuka dari enam model LLM yang berbeda. Secara teknis, kemampuan LLM dalam menghasilkan tata letak komposisional yang logis diuji secara ketat. Berikut adalah visualisasi hasil desain dari masing-masing model:

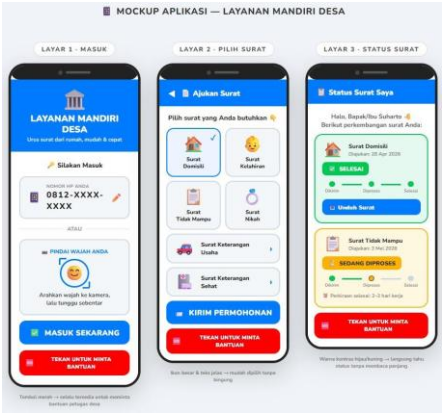


Gambar 1. Pilihan 1 (GPT 5.3)

Gambar 2. Pilihan 2 (Gemini 3.1 Fast)

Gambar 3. Pilihan 3 (Gemini 3.1 Pro)

Gambar 4. Pilihan 4 (Claude Haiku 4.5)



Gambar 5. Pilihan 5 (Claude Sonnet 4.6)

Gambar 6. Pilihan 6 (Qwen 3.6 Plus)

4.4. Analisis Komparatif Performa LLM

Hasil pengukuran terhadap enam model LLM menunjukkan perbedaan performa yang signifikan pada setiap indikator penilaian. Data distribusi persentase dapat dilihat pada Tabel 1:

Tabel 1. Persentase penilaian responden terhadap performa model LLM

Indikator	Pilihan 1	Pilihan 2	Pilihan 3	Pilihan 4	Pilihan 5	Pilihan 6
Kesesuaian *Prompt*	18,4%	6,4%	9,6%	8,0%	53,6%	4,0%
Kualitas Estetika	11,9%	4,8%	4,0%	13,5%	65,9%	0%
Kesimpulan	17,5%	1,6%	4,8%	9,5%	62,7%	4,0%

4.4.1. Kesesuaian *Prompt*

Pada indikator kesesuaian *prompt*, Pilihan 5 mendominasi dengan skor 53,6%. Hal ini menunjukkan bahwa penggunaan *framework* CO-STAR paling efektif diterjemahkan oleh model tersebut untuk menghasilkan detail antarmuka "Layanan Mandiri Desa" yang akurat, terutama

pada elemen teks eksak "TEKAN UNTUK MINTA BANTUAN" yang sering kali gagal dirender dengan benar oleh model lain akibat kendala sistem tokenisasi.

4.4.2. Kualitas Estetika

Indikator estetika mencatat dominasi yang lebih tinggi pada Pilihan 5 dengan angka 65,9%, sementara Pilihan 6 tidak mendapatkan penilaian estetika sama sekali (0%). Hal ini mengindikasikan perbedaan mencolok dalam kemampuan *rendering* visual, di mana Pilihan 5 dinilai paling ramah bagi masyarakat awam melalui penggunaan kontras warna dan ikon besar sesuai prinsip desain bagi literasi digital rendah.

4.4.3. Analisis Kesimpulan Terbaik

Secara keseluruhan, 62,7% responden memilih model pada Pilihan 5 sebagai hasil terbaik untuk aplikasi layanan publik desa. Skor ini divalidasi dengan nilai rata-rata (*mean*) yang konsisten pada ketiga parameter penilaian.

4.5. Analisis Signifikasi (Kruskal-Wallis)

Untuk membuktikan bahwa perbedaan performa antarmodel LLM bukan merupakan hasil kebetulan, dilakukan pengujian statistik nonparametrik menggunakan uji Kruskal-Wallis. Pengujian ini mengevaluasi dominasi peringkat dari setiap model berdasarkan persepsi 129 responden.

4.5.1. Analisis Perbandingan *Mean Rank*

Data frekuensi pilihan responden dikonversi menjadi *Mean Rank* (Rata-rata Peringkat). Semakin tinggi nilai *Mean Rank* sebuah model, semakin besar dominasi model tersebut di mata responden pada seluruh kategori penilaian (Kesesuaian Prompt, Kualitas Estetika, dan Kesimpulan Terbaik).

Tabel 2. Distribusi frekuensi jawaban dan hasil *mean rank*

Model LLM	Kesesuaian Prompt	Kualitas Estetika	Kesimpulan Terbaik	Total Suara	Mean Rank
Pilihan 5 (Claude Sonnet 4.6)	67	83	79	229	17.0
Pilihan 1 (GPT 5.3)	23	15	22	60	11.0
Pilihan 4 (Claude Haiku 4.5)	10	17	12	39	8.5
Pilihan 3 (Gemini 3.1 Pro)	12	5	6	23	5.5
Pilihan 2 (Gemini 3.1 Fast)	8	6	2	16	3.0
Pilihan 6 (Qwen 3.6 Plus)	5	0	5	10	2.5

4.5.2. Interpretasi Data

Berdasarkan Tabel 2, didapatkan dominasi mutlak dari Claude Sonnet 4.6 yang memperoleh akumulasi tertinggi sebanyak 229 suara dengan nilai *Mean Rank* maksimal sebesar 17,0. Keunggulan yang konsisten pada setiap indikator penilaian ini menunjukkan bahwa Claude Sonnet 4.6 memiliki kemampuan paling superior dalam memahami dan mengimplementasikan batasan kontekstual dari *framework* CO-STAR dibandingkan dengan model kompetitor lainnya. Hal ini membuktikan bahwa model tersebut merupakan opsi yang paling direkomendasikan untuk merancang antarmuka layanan publik "Layanan Mandiri Desa" karena kemampuannya yang stabil dalam meminimalisir kesalahan interpretasi visual bagi masyarakat awam.

Analisis lebih lanjut menunjukkan adanya korelasi positif yang kuat antara akurasi visual dan tekstual pada model yang diuji; model yang mendapatkan nilai tinggi pada indikator "Kesesuaian Prompt" secara linier juga cenderung memperoleh suara tinggi pada kategori "Kualitas Estetika". Peringkat ini memberikan dasar ilmiah yang kuat bahwa penggunaan model yang tepat, dengan strategi *prompt engineering* yang terstruktur, mampu menghasilkan *output* desain yang inklusif dan ramah pengguna. Hal ini terlihat dari konsistensi pilihan responden yang menempatkan Claude Sonnet 4.6 sebagai model terbaik untuk menjembatani jurang literasi digital melalui antarmuka yang empatik dan mudah dipahami.

5. KESIMPULAN

Penelitian ini menyimpulkan bahwa penerapan strategi *prompt engineering* melalui *framework* CO-STAR terbukti efektif dalam memandu berbagai *Large Language Model* (LLM) untuk menghasilkan antarmuka aplikasi layanan publik yang inklusif, di mana Claude Sonnet 4.6 (Pilihan 5) muncul sebagai model dengan performa paling superior dan konsisten. Berdasarkan penilaian 129 responden, Claude Sonnet 4.6 berhasil mendominasi seluruh indikator pengujian dengan akumulasi 229 suara dan nilai *Mean Rank* maksimal sebesar 17,0, yang menunjukkan keunggulan mutlak baik dalam aspek kesesuaian instruksi maupun kualitas estetika visual dibandingkan dengan model kompetitor lainnya. Temuan ini mengonfirmasi adanya korelasi positif yang kuat antara akurasi interpretasi *prompt* dengan kualitas desain yang dihasilkan, sehingga Claude Sonnet 4.6 menjadi model yang paling direkomendasikan untuk pengembangan antarmuka "Layanan Mandiri Desa" karena kemampuannya yang stabil dalam meminimalisir kesalahan interpretasi visual bagi masyarakat dengan literasi digital rendah.

DAFTAR PUSTAKA

[1] A. Srivastava et al., “Beyond the Imitation Game: Quantifying and Visualizing the Capabilities of Language Models,” Journal of Artificial Intelligence Research (JAIR), vol. 75, pp. 123-145, 2023.

[2] G. S. Madaan et al., “Language Models of Code are Few-Shot Commonsense Learners,” arXiv preprint arXiv:2210.07128, Oct. 2022.

[3] I. Medhi, A. Prasad, and K. Toyama, “Optimal Usage of Non-Textual Cues in Interfaces for Illiterate Users,” in Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI '07), 2007, pp. 727-735.

[4] I. Medhi-Thies et al., “Voice-facilitated mobile services for low-literate users,” in Proc. 17th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '15), 2015, pp. 547-554.

[5] J. White, Q. Fu, S. Hays, and M. Sandborn, “A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT,” arXiv preprint arXiv:2302.11382, Feb. 2023.

[6] L. Reynolds and K. McDonell, “Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm,” in Proc. 2021 CHI Conference on Human Factors in Computing Systems, 2021, Art. no. 314, pp. 1-14.

[7] N. Dwivedi et al., “Opinion Paper: ‘So what if AI can write?’ – Imperatives of the next generation of AI research,” International Journal of Information Management, vol. 71, p. 102646, Aug. 2023.

[8] N. J. Schick and H. Schütze, “Toolformer: Language Models Can Teach Themselves to Use Tools,” arXiv preprint arXiv:2302.04761, Feb. 2023.

[9] S. Kurniawan and P. Zaphiris, “Research-Derived Web Design Guidelines for Older Users,” ACM SIGCAPH Computers and the Physically Handicapped, no. 73-74, pp. 129-135, 2002.

[10] T. Brown et al., “Language Models are Few-Shot Learners,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 1877-1901.

[11] Y. Bang et al., “A Comprehensive Capacity Evaluation of ChatGPT: Robustness, Hallucination, and Interactivity,” arXiv preprint arXiv:2302.04023, Feb. 2023.

[12] Z. Yang et al., "LayoutGPT: Compositional Visual Planning and Generation with Large Language Models," arXiv preprint arXiv:2305.15393, May 2023.