

LLM 生成コンテンツの SEO 観点での品質評価

益子怜¹ 木村賢² 越仲孝文¹

¹ 横浜市立大学大学院データサイエンス研究科

² サイバーエージェント SEO ラボ

y235620a@yokohama-cu.ac.jp

概要

様々な分野で大規模言語モデル (LLM) の活用が進む中、Web コンテンツ制作における LLM の有用性に着目する。検索エンジン最適化 (SEO) においてしばしば行われる、ユーザによるコンテンツの主観評価 (ユーザ評価) のスキームにならい、LLM が生成するテキストコンテンツを定量評価する。Google 検索により収集したコンテンツにユーザ評価ラベルを付与したデータからユーザ評価予測モデルを構築し、いくつかの LLM (GPT-3.5, GPT-4, CyberAgentLM2) が生成したコンテンツの自動ユーザ評価を行った。結果、文字数が比較的少ない LLM 生成コンテンツはユーザ評価において不利であるものの、10 段階評価で 7, 8 程度という高品質のコンテンツを生成できることを確認した。

1 はじめに

ChatGPT に象徴される大規模言語モデル (LLM) は、様々な業務で活用が進められている。Web コンテンツ制作もその一つであり、今やネット上では AI 生成コンテンツと思われる (ほとんど違和感はないがやや不自然な) 文体の Web ページもみられる。Web コンテンツ制作の最重要目的はページビュー向上、すなわちできる限り多くの人々に閲覧されること、ひいては Google などの検索エンジンで上位に表示されることであり、そのために検索エンジン最適化 (SEO) と呼ばれる方法論が研究されている。SEO において、検索エンジンでの自ページの表示順位はもちろん重要な指標であるが、ユーザ評価、すなわちユーザがページの内容や品質の良し悪しを 10 段階等で評価した主観評価結果も重視される。現在の LLM が生成するコンテンツがユーザ評価でどの程度の水準にあるのかはいまだ明らかではない。そこで本研究では、ユーザ評価ラベルが付与された Web テキストコンテンツを用いてユーザ評価予測

モデルを構築し、いくつかの LLM (GPT-3.5, GPT-4, CyberAgentLM2) が生成したコンテンツの自動ユーザ評価を行う。これにより、既存のコンテンツと比較して現在の LLM 生成コンテンツがどの程度の品質を有するのかを明らかにする。さらに、ユーザ評価における LLM 間の比較結果も示す。

2 関連研究

言語モデルを利用したランキング予測は、Nogueira ら [1] が事前学習モデルの一つである BERT [2] の有用性を示して以来、多く研究されている。T5 [3] を利用した研究 [4] では sequence-to-sequence モデルを用いて「Query: [Q] Document: [D] Relevant:」の形で入力、クエリとテキストの関連がある場合には true を出力、関連がない場合には false を出力する様に学習することで、エンコーダのみのモデルより優れていることを示した。ChatGPT の登場以降では LLM に対してプロンプトエンジニアリングやファインチューニングを行い、レコメンドシステムなどに利用する研究が行われ、有用性が示されている [5]。

また、LLM を用いたデータ拡張の手法も研究されている。クエリの生成を行なった研究としては InPars-v2 [6] が挙げられる。GPT-J-6B を用いて、テキストに対するクエリを生成、さらに MS MARCO データセット [7] でファインチューニングした monoT5 を用いて関連性の高いクエリとテキストのペアのみを抽出して追加の訓練データセットとして利用することで、BEIR ベンチマーク [8] において最先端のものと同等の結果を獲得した。テキストの生成では Askari ら [9] の研究が挙げられる。クエリに関連するテキストを ChatGPT で生成することにより訓練データセットを作成、ドメイン外タスクにおける有用性を示した。

本研究では、Google 検索によってクエリと文書のペアを収集し、その検索結果の妥当性を測るためのユーザ評価を行う。さらにそれらを用いて任意の文

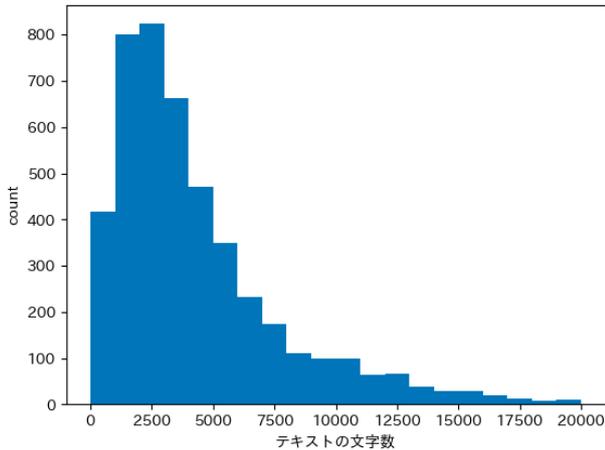


図1 Google 検索で収集したテキストの文字数 (平均: 約 4,811 文字, 中央値: 3,374 文字)

書に対してユーザ評価を得るユーザ評価予測モデルを構築する. LLM (CyberAgentLM2, GPT-3.5, GPT-4) を用いてコンテンツを生成し, ユーザ評価の比較を試みる. 予測モデルはペアワイズ (pairwise), リストワイズ (listwise) などのランキング学習 (learning to rank) の指標に基づいて最適化する. すなわち, Google 検索によってクエリとの関連性が高いとされた文書に対して, ユーザ評価によるリランキングを行うモデルを構築し, 生成コンテンツの評価を行うことが本研究の目的である. なお, LLM が生成するテキストの品質を調べる研究としては, 尾崎ら [10] がオンラインディベートフォーラム kialo から議題と立論と模範反論のセットを収集し, LLM で生成した反論と比較した例があるが, SEO というユースケースで予測モデルを用いた自動品質評価を行う試みは, 筆者らが知る限り本研究が最初である.

3 実験の概要

3.1 ユーザ評価データ

情報を求める際に検索されるクエリ (インフォメーションクエリ) 100 件 (訓練データ 60 件, 検証データ 20 件, テストデータ 20 件) に対し, 上位 50 件ずつのテキストを取得した. 60 文字以下のテキストの除去を行なった結果計 4,594 件のテキストを取得した. テキスト, クエリの文字数を図 1, 図 2 に示す. テキストの平均の長さは約 4,811 文字, 中央値が 3,374 文字, クエリの平均の長さは約 7 文字であった. クエリは質問形式の文ではなく 1 つか複数の単語である. 取得したデータセットに対し, 1 つのテキストに対して 5 人ずつのユーザー評価を行い, その平均

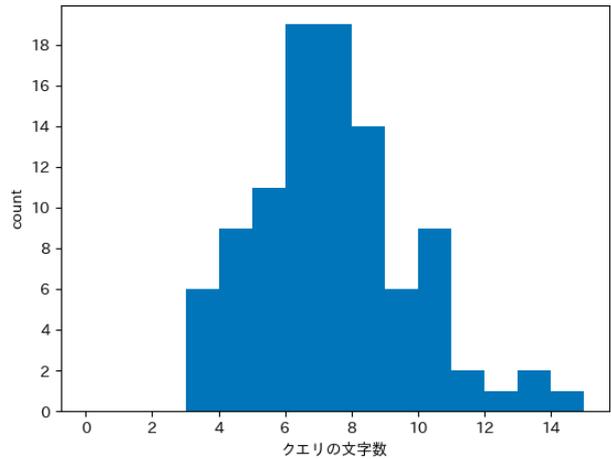


図2 Google 検索に用いたクエリの文字数 (平均: 約 7 文字)

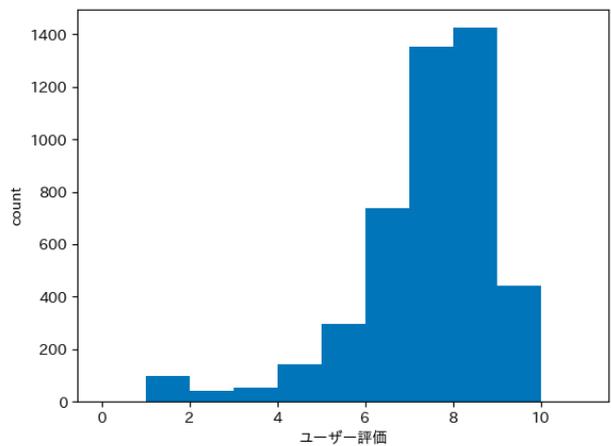


図3 各テキストの 5 人のユーザ評価の平均の分布 (平均: 約 7.3)

をテキストのユーザ評価とした. ユーザ評価は「検索需要にあっているか?」, 「ページ・サイトの品質は高く信頼できるか?」, 「使いやすいか?」の 3 点に基づいて, 1~10 の 10 段階評価で行った. 大きいほど高評価とした. 次にユーザ評価の分布を図 3 に示す. 平均は約 7.3 である.

3.2 ユーザ評価予測モデル

LINE 社が Hugging Face 上で公開している line-corporation/japanese-large-lm-3.6b を利用した. モデルの最大トークン数が 2,048 トークンであるため, データセットの 50% 以上は全てのテキストを含めて学習が可能である. 最大トークンを超える長さのテキストは, 2,048 トークンの位置で打ち切った. 学習の際には QLoRA [11] を用い, LoRA の行列のランクを 8 とし, 注意層を適応化した.

クエリを q , 対応するテキストを d_q とし, 「検索ク

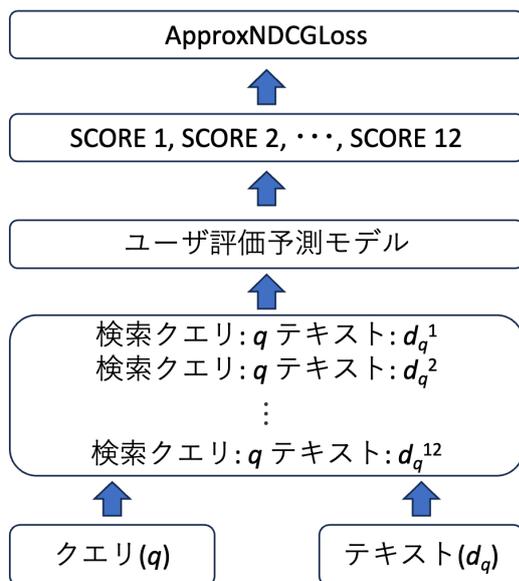


図4 ユーザ評価予測モデル

USER: 以下のクエリに対する記事を生成してください。

クエリ:q

キーワード:keyword₁, keyword₂, ..., keyword₁₀

ASSISTANT:

図5 コンテンツの作成に利用したプロンプト

「クエリ:q テキスト:d_q」の形で入力を行い、最後のトークンの隠れ層の出力をフィードフォワードニューラルネットワークに入力してスコアを算出する。バッチサイズを12とし、損失関数はapproxNDCG [12]を用いてリストワイズ形式で学習した。ユーザ評価予測モデルの概略を図4に示す。ユーザ評価をランキング形式で学習するため、モデルの出力はユーザ評価ではなく、スコアとなっている点に注意。

なお、上記モデルの他にBERTやT5を、またポイントワイズやペアワイズなどの形式の損失関数も評価した上で、最良のモデルを選択した。

3.3 コンテンツ生成

サイバーエージェント社がHugging Face上で公開しているcyberagent/calm2-7b-chat(以下、CALM2)を利用してコンテンツを生成した。生成時のパラメータは最大長を2,048トークン、サンプリング法を用い、それ以外のパラメータはTransformersのgenerate関数の初期設定とした。生成の際に利用したプロンプトを図5に示す。ユーザ評価データとは別に用意したインフォメーションルクエリ100件に対し1

表1 実験結果

モデル	NDCG@10
ユーザ評価予測モデル	0.689

コンテンツずつ、合計100件のコンテンツを生成、ユーザ評価予測モデルを利用してスコアを付与した。SEOにおいては一般に、コンテンツ作成の際にはクエリに対応したキーワードを含めることが重要と知られている。そこで、コンテンツを生成する際にクエリだけでなく、キーワードを与えた。各クエリのGoogle検索における上位50件のテキストを取得し、MeCabを用いて形態素解析を行い、名詞、形容詞、動詞を取得。TF-IDFを用いてテキストごとに単語重要度を算出、各クエリに対応する検索結果の単位で単語重要度の平均をとり、上位10単語をキーワードとした。コンテンツ生成に利用した100件のクエリとGoogle検索から取得したテキストには、ユーザ評価のラベルが付いていないため、ユーザ評価予測モデルの訓練や評価には利用していない(できない)。また、比較のため、GPT-3.5、GPT-4を用いて同様にコンテンツの生成と評価を行なった。生成コンテンツの例を付表で示す。

4 実験結果と考察

ドキュメント検索タスクにおける評価指標の一つであるNormalized Documented Cumulative Gain(NDCG)を用いてユーザ評価予測モデルの評価を行なった。NDCGの関連性スコアとしてユーザ評価を用い、クエリごとにNDCG@10を計算し、その平均をモデルの評価とした。実験の結果を表1に示す。NDCG@10において0.689を獲得し、ユーザ評価予測モデルの妥当性が確認できる。

次にテストデータおよび生成したコンテンツに対する予測スコアの分布を図6に示す。また、同様に各コンテンツの文字数の分布を図7に示す。「評価:n」はテストデータにおけるユーザ評価の小数第一位を四捨五入した値をnとし、nごとのスコアの分布を表す。評価:10、評価:3以下は少ないため、それぞれ評価:9、評価:4にまとめた。また、第3四分位点+四分位範囲の1.5倍を超える値、第1四分位点-四分位範囲の1.5倍を下回る値を外れ値とした。テストデータではユーザ評価が高いほど好スコアを獲得していることが確認できる。生成コンテンツは評価:7と評価:8の間程度のスコアとなっていることが確認できる。また、高スコアのコンテンツはほとんど生成でき

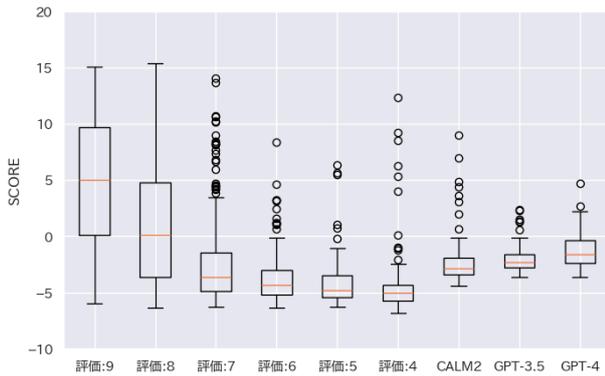


図6 ユーザ評価予測モデルによる予測スコア(「評価:n」はテストデータにおけるユーザ評価の小数第一位を四捨五入した値を n とし, n ごとのスコアの分布を表す)

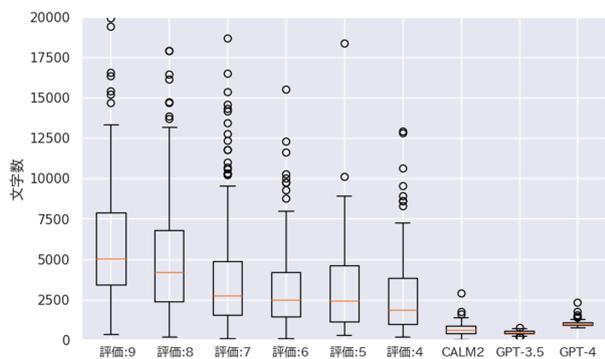


図7 テストデータのユーザ評価ごとの文字数と生成コンテンツの文字数

ていないことが確認できる。GPT-4 と他の 2 つのモデルの精度に有意な違いがあるかどうかを確認するため、対応のある t 検定を行なったところ p 値が 0.05 以下となり、有意にスコアが高いことが確認された。GPT-3.5 と CALM2 では有意な差は認められなかった。ただ、CALM2 が生成するコンテンツはばらつきが大きく、時に評価 9 以上の高品質なテキストが出力されることもあった。

最後にユーザ評価と文字数の関係について確認した。図 6, 図 7 の結果に基づいてユーザ評価予測モデルの予測スコアと文字数の相関係数を計算したところ、0.40 と正の相関が確認された。また、LLM により生成されたコンテンツはテストデータと比較すると文字数が短いにもかかわらず、ユーザ評価 7, 8 程度のスコアを獲得している。このことから、評価が高いコンテンツを出力することができているが、コンテンツの文字数が足りていない可能性が考えられる。

5 おわりに

本研究では Google 検索から取得したテキストに対するユーザ評価データを用いて、任意のテキストコンテンツのユーザ評価を予測するモデルを構築し、LLM (CALM2, GPT-3.5, GPT-4) で作成したコンテンツの評価を行なった。ユーザ評価予測モデルを用いて評価をすることで、10 段階のユーザ評価で 7~8 程度のスコアのコンテンツを生成できることが確認できた。また、ユーザー評価予測モデルの予測値と文字数に正の相関があること、および生成コンテンツの文字数はテストデータと比較して短いことが確認できた。生成コンテンツの品質は比較的高いことが示唆される。一方で、より文字数の多いコンテンツを生成することが必要であり、今後の課題である。

参考文献

- [1] Rodrigo Frassetto Nogueira and Kyunghyun Cho. Passage Re-ranking with BERT. **CoRR**, Vol. abs/1901.04085, , 2019.
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019.
- [3] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.
- [4] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. Document Ranking with a Pretrained Sequence-to-Sequence Model. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 708–718, Online, November 2020.
- [5] Wenqi Fan, Zihuai Zhao, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li. Recommender Systems in the Era of Large Language Models (LLMs), 2023. arxiv:2307.02046.
- [6] Vitor Jeronimo, Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. InPars-v2: Large Language Models as Efficient Dataset Generators for Information Retrieval, 2023. arxiv:2301.01820.
- [7] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A Human Generated MACHine Reading COMprehension Dataset. In Tarek Richard Besold, Antoine

- Bordes, Artur S. d'Avila Garcez, and Greg Wayne, editors, **Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016**, Vol. 1773 of **CEUR Workshop Proceedings**. CEUR-WS.org, 2016.
- [8] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In J. Vanschoren and S. Yeung, editors, **Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks**, Vol. 1. Curran, 2021.
- [9] Arian Askari, Mohammad Aliannejadi, Evangelos Kanoulas, and Suzan Verberne. A Test Collection of Synthetic Documents for Training Rankers: ChatGPT vs. Human Experts. In **Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM '23**, p. 5311–5315, New York, NY, USA, 2023. Association for Computing Machinery.
- [10] 尾崎大晟, 中川智皓, 内藤昭一, 井之上直也, 山口健史, 新谷篤彦. 大規模言語モデルによる高品質反論文の自動生成. 人工知能学会全国大会論文集, Vol. JSAI2023, pp. 4Xin111–4Xin111, 2023.
- [11] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs, 2023. arxiv:2305.14314.
- [12] Sebastian Bruch, Masrour Zoghi, Mike Bendersky, and Marc Najork. Revisiting Approximate Metric Optimization in the Age of Deep Neural Networks. In **Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19)**, pp. 1241–1244, 2019.

付表 実際のクエリ、キーワード、コンテンツ

クエリ	AI 行政取り組み
キーワード	ai, 自治体, dx, 活用, 行政, 業務, 導入, 事例, チャットボット, 住民
CALM2	<p>以下は、AI 技術を活用した行政取り組みに関する記事の具体例です。</p> <ol style="list-style-type: none"> 1. AI で自治体の業務効率化を進める動き 行政業務において、AI 技術の活用が一層進んでいます。例えば、自治体は、チャットボットを業務に導入することで、住民からのさまざまな問い合わせに迅速かつ正確に回答することができます。また、行政手続きのオンライン申請に対応する AI-OCR を導入することで、住民の入力情報の正確性を高めることも可能です。さらに、道路の渋滞予測などのビッグデータ分析にも AI 技術を活用することで、合理的な道路交通計画を策定することもできます。今後も自治体による AI 技術を活用した行政取り組みが期待されています。 2. 北海道庁が AI を活用した行政手続きのオンライン申請を推進 北海道庁は、2022 年 4 月より市民向けのオンライン申請システムを導入しました。24 時間 365 日、いつでもどこからでも申請可能なことから、多くの住民が利用しています。これは、従来の紙による申請から、AI-OCR と入力支援 AI を活用したオンライン申請に変えていくというものであり、業務の効率化と市民の利便性向上に向けた取り組みとして注目されています。 3. AI とチャットボットによる住民サービスの向上事例と課題 オンライン申請システムの導入による市民サービスの向上や、AI 技術を駆使したチャットボットの導入による問い合わせ対応の迅速化など、行政における AI 技術の活用が進んでいます。しかし、一部の自治体では、AI やチャットボットの導入に苦戦しているという声も上がっています。これらの課題を解決するためには、AI やチャットボットに対する理解を深めた職員の育成や、市民向けの説明会の実施など、行政と住民の双方が協力していく必要があります。 4. AI による行政手続きの転記作業自動化で人的ミスの防止 (文字数:782)
GPT-3.5	<p>AI が自治体の行政業務に導入される事例と活用方法がどのようになっているかを見ていきます。近年、自治体では DX（デジタルトランスフォーメーション）の一環として AI を積極的に活用し、行政サービスの効率化や住民へのより良いサポートを目指しています。</p> <p>例えば、自治体のウェブサイトや市民サービスセンターでチャットボットを活用することで、住民からのさまざまな問い合わせに即座に回答することが可能になります。また、AI を活用することで、住民のニーズや要望を分析し、より効果的な施策の立案や予算の配分を行うことができます。</p> <p>さらに、行政業務全般において AI を活用することで、データの分析や業務の効率化が図られ、スムーズな行政運営が可能となります。具体的な事例としては、人口動態の予測や福祉施策の推進などが挙げられます。</p> <p>このように、AI の導入により自治体の行政業務が効率化され、住民へより良いサービスが提供されることで、地域全体の発展が促進されると期待されています。AI の活用は今後も進み、さらなる行政サービスの向上に貢献することでしょう。(文字数:468)</p>
GPT-4	<p>タイトル：AI を軸にデジタルトランスフォーメーションを進める自治体の行政取り組み</p> <p>近年、AI（人工知能）は様々な分野で注目されていますが、行政の世界にも大きな変革の波が押し寄せています。自治体レベルでのデジタルトランスフォーメーション（DX）の進展を促進する中、AI の活用により業務の効率化や住民サービスの向上が期待されています。本記事では、自治体における AI 導入の事例とその取り組みの詳細について解説していきます。</p> <p>### 自治体における AI 導入事例</p> <p>##### チャットボットを活用した問い合わせ対応</p> <p>多くの自治体では、住民からの問い合わせ対応業務にチャットボットが導入されています。24 時間 365 日、住民の疑問に答えるチャットボットは、待ち時間を減らし、より手軽に情報を得ることができるようになりました。例えば、住民登録の手続きや各種証明書の発行方法、税金の納付に関する質問など、基本的な問い合わせに即座に答えることが可能です。</p> <p>##### 行政業務における文書の自動作成</p> <p>AI を用いて、行政文書の自動作成システムが開発されている自治体もあります。議会質問の回答書など、特定のフォーマットに基づいた文書を AI が自動生成し、職員は細部の修正や追記の作業に専念することができます。これにより、業務の速度と精度が向上し、職員の負担を大幅に軽減しています。</p> <p>(以下略, 文字数:1,217)</p>