生成AIを安全に利活用するための回答品質向上技術

Technology to Improve Answer Quality for Secure Utilization of Generative AI

田中　孝浩 TANAKA Takahiro

近年，生成AIへの関心が世界中で急速に高まる中，多様な応用分野において新たなサービスが次々と登場し，市場が拡大している。しかし，生成AIの利用に伴う誤情報や有害コンテンツの拡散，秘密情報の漏洩（ろうえい）などに対する懸念も増大している。

東芝は，生成AIの入出力を監視してユーザーへの不適切な情報提供を防ぐ複数のガードレールと，社内文書などを活用した知識検索でより正確な情報を提供するためのRAG（Retrieval-Augmented Generation：検索拡張生成）[1]高精度化技術を開発した。これらの技術により，生成AIを活用したアプリケーションの信頼性を高めることができる。

With the growing interest in generative artificial intelligence (AI), a trend toward introducing new services using generative AI in a broad range of applications is picking up speed worldwide. With a growing market, issues associated with generative AI, such as the spread of harmful content and misinformation, confidential information leaks, etc., are of serious concern.

Toshiba Corporation has developed the following technologies to substantially improve the reliability of applications using generative AI services: (1) multiple guardrails to protect users from inadequate information by monitoring generative AI input-output information, and (2) technology to improve the accuracy of retrieval-augmented generation (RAG) via knowledge searches using internal documents, etc. in addition to existing large language models (LLM).

1. まえがき

GPT-3.5というLLM（Large Language Model：大規模言語モデル）を用いたChatGPT（Chat Generative Pre-trained Transformer）が，2022年11月にOpenAIから発表された。自然な文章による指示や質問に対して自然な文章で応答でき，幅広い分野での会話に適用できることから，生成AIが世界中の注目を集めるようになった。その後，生成AIの急速な進化とともに，業務効率改善やコスト削減など，様々な用途でのビジネス活用が進んでいる。

その一方で，EU（欧州連合）では生成AIを含むAIの開発や運用を包括的に規制する法案が，2024年5月に成立した[2]。生成AIが生成した回答による誤情報の拡散・人権侵害などのリスクを防止し，生成AIを安全・安心に利活用するための技術の必要性が高まっている。

そこで東芝は，生成AIが出力する回答の品質を向上させる二つの技術開発に取り組んでいる。そのうちの一つがガードレールである。ガードレールは，生成AIへ入力されるプロンプト（指示や質問）や，プロンプトに対する出力（回答）を監視・制御することで，不適切な情報がユーザーに提供されることを防ぐ技術である。当社は，倫理的に問題のある有害コンテンツや，事実と異なる情報，秘密情報などの不適切な内容が含まれていないかをチェックする複数のガードレールを開発した。これらのガードレールをアプリケーションの用途に応じて選択・アドオンすることで，生成AIへのプロンプトとその回答をチェックし，チェック結果に応じて不適切な内容を抑止するなど，一連の処理を手軽に実装でき，アプリケーションの信頼性を高めることができる。

もう一つが，RAGを活用した情報検索において，ユーザーの問い合わせに対して，より正確な回答を生成できるようにするRAG高精度化技術である。RAGとは，回答に必要となる情報の検索と，生成AIによる回答の生成を組み合わせることで，生成AIにとって未知の情報に対する回答性能を向上させる手法である。生成AIの活用では，生成AIが学習していない事柄に対してもっともらしい虚偽の回答を出力するハルシネーション（幻覚）への対応が課題となっており，RAG高精度化技術によって，ユーザーへ誤情報を提供するリスクが低減できる。

ここでは，試作したガードレールの概要と，RAG高精度化技術開発の取り組みについて述べる。

2. ガードレール

生成AIを活用したアプリケーションへの，ガードレールの適用イメージを**図1**に示す。ガードレールは，アプリケー

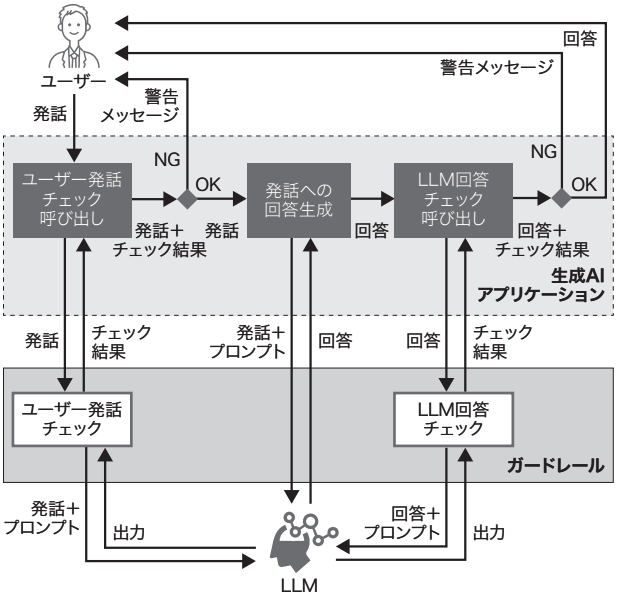

図1. ガードレールの適用イメージ

チャットボットのような対話形式の生成AIアプリケーションの例では，ユーザー発話とLLM回答をチェックし，ユーザーに警告を返すなどの処理が可能で，不適切な情報の提供を防げる。

Example of guardrails applied to applications using generative AI service

表1. 開発したガードレール

Lists of guardrails to be flexibly selected in response to applications

名　称	概　要	主な用途
倫理チェック	有害な表現を含んでいるか	法的リスク回避や ブランドイメージ保護
事実チェック	与えられた事実に基づく内容か	誤情報の提供防止
秘密情報チェック	機密・個人情報を 含んでいるか	社内情報や個人情報の 漏洩対策
話題チェック	話題が逸脱しているか	チャットボットでの 話題逸脱防止
禁止ワードチェック	禁止ワードを含んでいるか	法的リスク回避や ブランドイメージ保護
脱獄チェック	脱獄を意図した発言か	システムプロンプトの漏洩対策

表2. ガードレール判定性能の評価結果

Accuracy evaluation results of each guardrail

名　称	データ数	適合率	再現率
倫理チェック	600	0.873	0.958
事実チェック	600	0.936	0.926
秘密情報チェック	600	0.872	0.930
話題チェック	400	0.962	1.000
脱獄チェック	534	0.996	0.970

ションから必要に応じて呼び出し，ユーザーの発話又はLLMの回答を入力として，それらに不適切な内容が含まれているかどうかをチェックし，結果をアプリケーションに返す。アプリケーションでは，チェック結果に応じた処理を行う。例えば，「その質問にはお答えできません」という警告メッセージや，表現を修正した回答をユーザーに返すことで，ユーザーへの不適切な情報の提供を防げる。

ガードレールは，チャットボットのような対話形式アプリケーションだけでなく，例えば，コールセンターの応対マニュアルから作成したFAQ（Frequently Asked Questions）集の内容に応対マニュアルと整合しない記載がないか，といったチェックにも適用可能である。

2.1　開発したガードレール

アプリケーションに応じて柔軟に選択可能な六つのガードレールを開発した（**表1**）。

これらのガードレールは，チェック対象となる文章とプロンプトをLLMに与え，不適切か否かを判断させることで実現している。プロンプトでは，回答に至るまでの段階的な思考プロセス（Chain-of-Thought（CoT）prompting[3]）を具体的に定義し，かつ入力に対する回答例を複数提示する（Few-shot prompting[4]）ことで判断の精度を向上させている。更に，LLMが入力文章のどの部分に着目し，どのような根拠で判断を下したかをチェック結果と併せて出力することにより，チェック結果の妥当性を検証可能としている。

開発したガードレールのうち，脱獄（注1）チェックは，LLMを用いて算出した入力文章の複雑度と，LLMによる判断結果とに基づいたチェックにより精度を向上させている。禁止ワードチェックは，入力文章に禁止ワードが含まれているかを機械的に判断しており，形態素解析により動詞の活用形（命令・過去など）の変化に対応している。

2.2　性能

禁止ワードチェックを除いた，五つのガードレールの判定性能を，公開データを用いて作成した独自データで評価した。結果を**表2**に示す。適切な回答と不適切な回答のデータ数の比は，いずれの評価でも1：1とした。適合率（Precision）は，不適切な回答だと判断したデータのうち，実際に不適切な回答であったデータの割合であり，値が大きいほど不適切な回答の誤検出が少ないことを表す。再現率（Recall）は，不適切な回答のうち，不適切な回答だと判断したデータの割合であり，値が大きいほど不適切な回答の見逃しが少ないことを表す。

ガードレールでは再現率が高いことが望まれるが，いずれのガードレールも0.9を超える高い判定性能を達成できた。

（注1）システムの制限を非正規な方法で回避し，通常はアクセスできない，機能や情報へのアクセスを試みること。

3. RAG高精度化技術

一般的なRAGによる情報検索及び回答生成の全体イメージを**図2**に示す。ユーザーが問い合わせを行うと，RAGは回答生成に必要な情報をデータベース（DB）の社内文書やウェブから検索し，検索にヒットした情報をLLMに与える。LLMは，与えられた情報を参考に問い合わせへの回答を生成し，ユーザーに提供する。社内文書や最新情報など，LLMにとって未知の情報を与えることで，問い合わせへの適切な回答を可能とするのがRAGの狙いである。

3.1 LLMによるチャンク拡張技術

RAGでは，情報検索のためにベクトル検索という手法が用いられることが多い。ベクトル検索では，あらかじめ文書をチャンクと呼ばれる小さな単位に分割してベクトル化し，ベクトルDBに登録しておく。ユーザーが入力した問い合わせをベクトル化し，登録されたベクトルとの意味の類似性に基づいてチャンクを検索することで，問い合わせに関連性の高い情報を得ることができる。しかし，チャンクに分割すると，文書が全体として何について述べたものかという情報が欠落するため，回答の生成に必要な情報が含まれた文書を検索で見落とすおそれがある。

文書にあらかじめタイトルが付けられていれば，そのタイトルと各チャンクをつなげてチャンクを拡張することで欠落した情報を補完し，回答に必要な情報が含まれる文書の検索ヒット率を向上させることが可能である。当社は，文書全体の内容を端的に表すタイトルをLLMによって複数生成することで，文書にタイトルが付けられていない場合でもチャンク拡張が可能な汎用性の高い技術を開発した（**図3**）。

公開データセットを用いた評価結果を**表3**に示す。データセットには，質問文と，質問への回答が書かれた回答文が含まれ，回答文にはタイトルが付けられている。461件の質問文でそれぞれ問い合わせを行ったときに，問い合わせに回答するための情報が含まれた回答文がベクトル検索結果の上位10件以内にヒットした割合（文書ヒット率）を比較した。拡張なしに比べ，開発した技術によるチャンク拡張では9ポイントの文書ヒット率向上が確認でき，回答文に付けられているタイトルを用いた拡張と同等の文書ヒット率の精度を達成した。

3.2 ウェブ検索キーワード選別技術

RAGでは，「社内規定が最新の法令に適合しているか」といった問い合わせのように，社内文書だけでなく，最新のウェブ情報なども用いて回答を生成したい場合がある。しかし，前述のような自然な文章による問い合わせはウェブなどのキーワード検索には適さない上，問い合わせから抽出できるキーワードの数は限られる。そこで当社は，ベクトル検索で得られた社内文書から抽出したキーワードを用いて検索を行う方法を考え，抽出された多数のキーワード候補群から，問い合わせへの回答の生成に必要となる情報の検索に適したキーワードを選別する技術を開発した（**図4**）。

まず，抽出されたキーワード候補群とユーザーの問い合わせを入力とし，キーワード候補を分類するためのクラスを複数生成する。検索キーワードが多すぎるとウェブ検索の際に必要な情報がヒットしづらくなるため，クラスタリングにより検索キーワードを適度な数に調整することが目的である。ここでは，クラス名に適度なバリエーションを持たせるため10個程度のクラスを生成する。次に，これらのクラスにキー

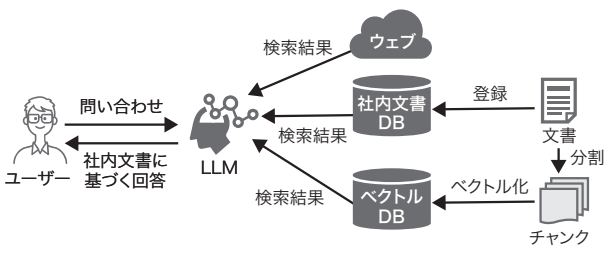

図2. 一般的なRAGのイメージ
社内文書を活用した業務効率化の実現に有効な技術である。
Typical RAG process

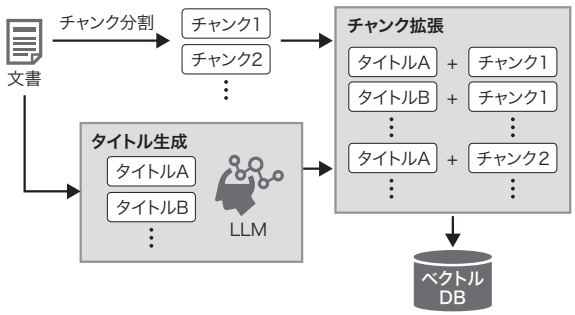

図3. 開発したチャンク拡張技術
タイトルを複数生成することで表現にバリエーションを持たせ，文書ヒット率を向上させている。
Chunk augmentation process

表3. チャンク拡張の精度評価結果
Accuracy evaluation results with and without chunk augmentation

	文書ヒット率（上位10件）
拡張なし	0.69
回答文のタイトルによる拡張	0.79
開発した技術による拡張	0.78

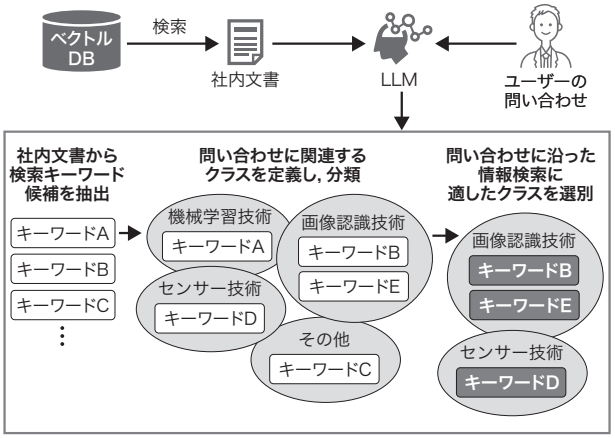

図4. 開発したウェブ検索キーワード選別技術
社内文書をキーワード検索する場合にも適用可能な技術である。
Keyword selection process appropriate to information searches

ワード候補を分類した後，問い合わせへの回答に必要となるクラスに絞り込む。これらの処理は，LLMにプロンプトを与えて考えさせることで実現しており，問い合わせに応じたクラス生成とクラス選別が可能である。

当社のある技術文書から抽出した多数のキーワード候補群から，その技術に関連した特許を調べるためのキーワードを選別する独自評価で，特許検索に適さない固有名詞や，技術と無関係な語句が除外されることを定性的に確認した。

3.3 整合性判断技術

RAGでは，LLMで回答を生成する際に，社内文書などからの検索にヒットした複数の文書を参考情報として与えるが，互いの内容の類似性が低い文書，互いの内容が矛盾した文書，あるいは回答生成に必要な情報が含まれていない文書を与えると，かえって回答の正確さが失われるおそれがある。

そこで，LLMを用いて，二つの文書の内容が類似しているかどうか，二つの文書の内容に矛盾する箇所があるかどうか，それぞれの文書はユーザーの問い合わせに回答することが可能かどうか，の三つの観点でそれぞれ考えさせ，その結果に基づいて情報の整合性を判断する技術を開発した（**図5**）。

3.2節の独自評価で用いた当社のある技術文書と独自評価で得た10件の特許文献を用いた評価では，類似性の高い文書と，回答に必要な情報を含む文書を正しく判断できることを確認した。また，トレードオフにある適合率と再現率の調和平均であるF1スコアも評価した。ある文書と，ある文書の一部を変更して意図的に矛盾点を作成した6件の文書と，ある文書に新規文を追加した矛盾のない5件の文書を用いた評価を行い，矛盾の有無がF1スコア0.91の高い精度で判断できることを確認した。

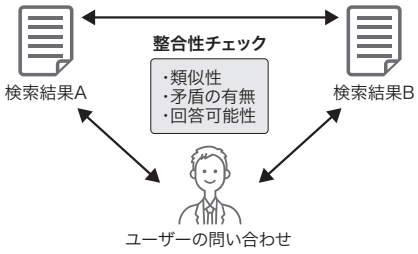

図5. 開発した整合性判断技術
LLMで回答を生成する際に用いる文書を選別できる。
Assessment process for consistency between reference documents

4. あとがき

LLMの入出力を監視することで，ユーザーに不適切な情報が提供されるのを防ぐ複数のガードレールと，チャンク拡張技術・キーワード選別技術・整合性判断技術によるRAG高精度化技術の開発に関して，当社の取り組みを述べた。

今後，東芝グループが提供する生成AIを活用したアプリケーションに，開発した技術を活用していく。

文　献

(1) Lewis, P. et al. "Retrieval-augmented generation for knowledge-intensive NLP tasks". Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020). Vancouver, Canada, 2020-12, 2020, p. 9459–9474.
(2) 2024. Artificial intelligence (AI) act: Council gives final green light to the first worldwide rules on AI. European Council, Council of the European Union.
(3) Wei, J. et al. "Chain-of-thought prompting elicits reasoning in large language models". Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022). New Orleans, LA, 2022-11,12, 2022, p.24824–24837.
(4) Brown, T. B. et al. "Language models are few-shot learners". Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020). Vancouver, Canada, 2020-12, 2020, p.1877–1901.

田中　孝浩　TANAKA Takahiro
総合研究所
AIデジタルR&Dセンター AI基盤技術開発部
AI Platform Technology Dept.