# 文章生成 AI が生成した家庭内危険行動の理由に対する根拠提示システム

A System to Show Evidence for Reason of Dangerous Behaviors in Home Generated by Sentence Generation AI

穴口史将 [*1]　森田武史 [*1*2]
Fumikatsu Anaguchi　Takeshi Morita

[*1]青山学院大学理工学部
College of Science and Engineering, Aoyama Gakuin University

[*2]産業技術総合研究所
National Institute of Advanced Industrial Science and Technology

To use sentence generation AI safely and securely, it is necessary to show evidence citing the literatures. Therefore, in this work, we propose a system to show evidence for reason of dangerous behaviors in home generated by sentence generation AI for the dataset presented by Knowledge Graph Reasoning Challenge 2023. First, we extract dangerous behaviors in home. Second, Sentence Generation AI generate reasons of it. Third, Retrieval Augmented Generation (RAG) retrieves a sentence similar to this reason from the literatures on dangerous behaviors in Home and shows the user as evidence. We did a survey to evaluate whether the sentence generation AI can appropriately generate reasons for dangerous behaviors in the home, and whether the evidence showed by the proposed system for these reasons is appropriate. As a result, average score of five-stage evaluation were 3.6 and 2.6. It is found that the proposed system can show general evidence for reasons.

## 1. はじめに

近年の大規模言語モデル（LLM: Large Language Model）に基づく文章生成 AI への関心の高まりに伴い，文章生成 AI が普及し様々な社会システムに利用されることが予想される．LLM は，ブラックボックス型のモデルであり，学習データから尤もらしい出力を生成している．そのため，その出力は正確性や信頼性が保証されず，根拠となる情報が暗黙的であり，再現性が担保されないという特徴を持つ．文章生成 AI 技術を安全・安心に社会の中で活用していくためには，システムが判断に至った理由とその根拠を，信頼性が高い文献を引用しながら説明する技術が必要になる．

このような背景のもと，人工知能学会セマンティックウェブとオントロジー研究会は，説明可能性を有する AI 技術の開発・促進を目的としたコンテスト「ナレッジグラフ推論チャレンジ」を開催している．先行研究として，「ナレッジグラフ推論チャレンジ【実社会版 2022】[鵜飼 22]」で公開されたデータセットから検出された家庭内危険行動を文章生成 AI の入力として，その行動が危険な理由を生成する研究 [浅野 23] がある．[浅野 23] で提案されたシステムは，尤もらしい出力を生成しているが，出力の正確性は保証されておらず，根拠となる文献などを提示できていないという課題があった．

以上より，本研究では，文章生成 AI が生成した家庭内危険行動の理由に対して，信頼性が高い文献を元に，根拠を提示可能なシステムを開発することを目的とする．

## 2. 関連研究

VirtualHome2KG[Egami 23] は，仮想環境で家庭内を再現した VirtualHome の情報をもとに構築した知識グラフ（KG: Knowledge Graph）であり，オープンデータとして公開されている．環境内に存在するオブジェクトに関する情報をノードとして，部屋や物に関する関係性をエッジとして定義している．また，ノードには，位置座標や状態などが定義されている．「ソファでくつろぐ」や「リビングルームの掃除」などの日常生活における行動（アクティビティ）の履歴を KG として構築する．アクティビティは，オブジェクトに対する、「座る」や「見る」などの複数の動作（アクション）から構築される．VirtualHome2KG は本研究のデータセットとして，提案システムや評価実験に使用する．

「ナレッジグラフ推論チャレンジ【実社会版 2022】」の参考アプローチとして，家庭内危険行動を検出する研究 [江上 22] や，応募作品として，家庭内危険行動の理由を生成する研究 [浅野 23] がある．[江上 22] では，VirtualHome2KG を用いて，家庭内事故の報告件数の中で上位二つの高齢者の転倒・転落を対象として，高齢者の家庭内危険行動を推論によって検出する手法を提案している．[浅野 23] では，検出された家庭内危険行動を，文章生成 AI が家庭内危険行動の理由を生成する手法を提案している．

本研究における家庭内危険行動の検出と理由の生成は，[江上 22] と [浅野 23] の研究を参考にした．

## 3. 提案システム

### 3.1 提案システムの構成

提案システムの概要を図 1 に示す．まず，[江上 22] を参考に，VirtualHome2KG から SPARQL クエリを用いて，高齢者の家庭内危険行動を検出し，家庭内危険行動 KG を構築する．次に，[浅野 23] を参考に，家庭内危険行動 KG から文章生成 AI を用いて家庭内危険行動の理由を生成する．次に，家庭内危険行動の理由に対して，検索拡張生成（RAG: Retrieval Augmented Generation）[Lewis 21] と Elasticsearch を用いて専門文書から類似文を検索し，根拠としてユーザに提示する．最後に，家庭内危険行動の理由に対する根拠の妥当性を検証する．

連絡先: 森田武史，青山学院大学理工学部，〒 252-5258 神奈川県相模原市中央区淵野辺 5-10-1，E-mail: morita@it.aoyama.ac.jp

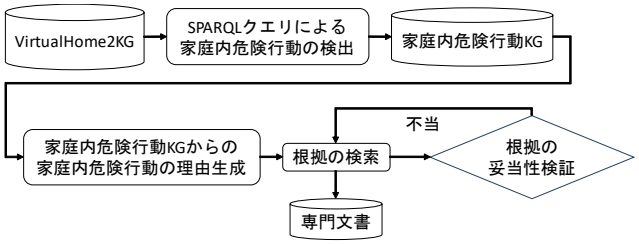


図 1: 提案システムの概要図

```
「{家庭内危険行動の理由}」
に最も関連する記述を1文で文書中から抜き出してください
最も関連する記述のみ出力てください．文書中にない記述を出力するのはやめてください
記述が無い場合は，'No Data'と出力してください
```

Listing 1: 根拠文を検索するプロンプトテンプレート

### 3.2 SPARQL クエリによる家庭内危険行動の検出

本研究では,「段差を上り下りする」,「高い場所にある物に手を伸ばす」,「低い棚から物を取り出す」の三つの家庭内危険行動を検出するための SPARQL クエリを作成し，これらに関する家庭内危険行動 KG を構築する．

### 3.3 家庭内危険行動 KG からの家庭内危険行動の理由生成

まず，文章生成 AI による家庭内危険行動の理由生成に必要な VirtualHome の環境情報を抽出する．二つのオブジェクト間の位置関係や，部屋とオブジェクトの包含関係などを SPARQL クエリを用いて抽出し，LLM を用いて部屋ごとに分類する．これら環境情報をプロンプトに含めることで，環境に基づいた家庭内危険行動を LLM が推論できる．次に，家庭内危険行動 KG から SPARQL クエリを用いて，アクティビティ，オブジェクト，アクションを抽出し，環境情報と共に，家庭内危険行動の理由を生成するプロンプトに組み込む．最後に，このプロンプトをもとに文章生成 AI が転倒・転落に関する家庭内危険行動の理由を生成する．

### 3.4 根拠の検索

3.3 節で述べた手法により生成された家庭内危険行動の理由に対して，RAG と Elasticsearch を用いて，その根拠を検索する．

まず，OpenAI 社の OpenAI Embeddings*1 を用いて，PDF ファイル形式の専門文書をベクトル化し，専門文書ベクトル DB に保存する．次に，LangChain の RetrievalQA を用いて，専門文書ベクトル DB から家庭内危険行動の理由に最も類似する記述を根拠として専門文書中から出力するように，文章生成 AI にプロンプトを与える．根拠文を検索するプロンプトテンプレートを Listing 1 に示す．Listing 1 の変数 “{ 家庭内危険行動の理由 }” に，3.3 節で述べた手法により生成された家庭内危険行動の理由を代入し，プロンプトを生成する．最後に，Elasticsearch を用いて根拠となる記述を専門文書から全文検索し，検索結果とそのメタデータを，家庭内危険行動の理由とその根拠と共に出力する．

*1 https://platform.openai.com/docs/guides/embeddings

```
危険な理由：テーブルの上に立っている際にバランスを崩し、転倒してしまう可能性があるから
根拠文：階段などの段差でつまずく、足がもつれて家具にぶつかる、ベッドから降りるときに転落する、靴下が引っかかって転落する、バスマットやじゅうたん、毛布などに足をとられて転倒するなど、ちょっとしたことが転落・転倒の原因になっている。
原文：階段などの段差でつまずく、足がもつれて家具にぶつかる、ベッドから降りるときに転落する、靴下が引っかかって転落する、バスマットやじゅうたん、毛布などに足をとられて転倒するなど、ちょっとしたことが転落・転倒の原因になっている
タイトル：医療機関ネットワークか事業からみた家庭内事故–高齢者編–
発行日：平成25年3月28日
著者：独立行政法人国民生活センター
ページ：3
```

Listing 2: 家庭内危険行動の理由に対する根拠の検索結果例

本研究で用いる専門文書のデータ量は比較的軽量かつ，複雑な処理を行う必要はないため，LangChain の RetrievalQA におけるパラメータのうち，chain_type には，ベクトル DB を単純に検索する “stuff” を使用する．

提案システムにおける Elasticsearch のインデックスは，根拠文，専門文書のタイトル，発行日，著者，ページである．データ型はそれぞれ，“text”，“text”，“text”，“text”，“integer” である．発行日には，年月日などの文字もデータに含めるため，テキスト型として作成した．インデックスは専門文書ごとに作成する．インデックスの構成要素である，ドキュメント（レコード）は文ごとに作成する．これによって一文ごとにメタデータを付与することができる．

家庭内危険行動の理由に対する根拠の検索結果例を Listing 2 に示す．Listing 2 の「危険な理由」は，文章生成 AI が生成した家庭内危険行動の理由である.「根拠文」は，LangChain の RetrievalQA を用いて専門文書ベクトル DB から検索した類似文である.「原文」,「タイトル」,「発行日」,「著者」,「ページ」は，Elasticsearch の検索結果である．

### 3.5 根拠の妥当性検証

3.4 節で述べた手法により検索された根拠文が，家庭内危険行動の理由に対する根拠として妥当であるかを検証する．検証後に根拠が不当であると判断された場合，再検索を行い，妥当な根拠を検索する．根拠の妥当性の検証を自動的に行うことは難しいため，本研究では手動で検証する．

## 4. 評価実験

### 4.1 評価実験概要

本評価実験では，文章生成 AI が家庭内危険行動の理由を適切に生成できるか，また，文章生成 AI が生成した家庭内危険行動の理由に対する根拠文が適切であるか，アンケートを用いて評価する．また，RAG と Elasticsearch が，それぞれ専門文書の文とメタデータを正確に出力できているかを評価する．

### 4.2 評価用データセット

評価用データセットには，VirtualHome2KG から SPARQL クエリにより検出した高齢者の家庭内危険行動を使用する．部屋数とアクション数が網羅的になるように，九つの家庭内危険行動を選択した．表 1 に評価用データセットから抽出した

表 1: 評価用データセット（三つ組）

| 識別子 | アクティビティ | オブジェクト | アクション |
|---|---|---|---|
| R1 | テーブルの上に立つ | テーブル | のぼる |
| R2 | 歯磨きの準備 | 歯ブラシ | 掴む |
| R3 | 冷蔵庫に食料品をしまう | 冷蔵庫 | 開ける |
| R4 | クローゼットの整理整頓 | シャツ | 置く |
| R5 | ソファでくつろぐ | ソファ | 座る |
| R6 | リビングルームの掃除 | テレビ | 拭く |
| R7 | テレビを見ながらジュースを飲む | テレビ | スイッチオン |
| R8 | 寝る | ライトスイッチ | スイッチオフ |
| R9 | コンロを使ってジャガイモを調理する | フライパン | 油を注ぐ |

表 2: 評価用データセット（文章化）

| 識別子 | 家庭内危険行動 |
|---|---|
| R1 | テーブルの上に立つためにテーブルにのぼる |
| R2 | 歯磨きの準備をするために歯ブラシを掴む |
| R3 | 冷蔵庫に食料品をしまうために冷蔵庫を開ける |
| R4 | クローゼットの整理整頓ためにシャツを置く |
| R5 | ソファでくつろぐためにソファに座る |
| R6 | リビングルームの掃除ためにテレビを拭く |
| R7 | テレビを見ながらジュースを飲むためにテレビのスイッチを押す |
| R8 | 寝るためにライトのスイッチを押す |
| R9 | コンロを使ってジャガイモを調理するためにフライパンに油を注ぐ |

家庭内危険行動の三つ組（アクティビティ，オブジェクト，アクション）を示す．また，アンケート用に，表 1 に示す三つ組を「アクティビティのために，オブジェクトにアクションする」といった文章に手作業で変換した結果を表 2 に示す．表 1 と表 2 の 1 列目には，家庭内危険行動の識別子として，R1 から R9 を付与する．評価実験結果では，1 列目の識別子を用いて説明する．

### 4.3 評価方法

提案システムを用いて，評価用データセットの家庭内危険行動の理由を生成し，家庭内危険行動の理由に対する根拠文を専門文書から検索する．使用する LLM は,「gpt-3.5-turbo-1106」である．また，本評価実験では，専門文書中に根拠文が存在することを前提とする．これらを踏まえ，以下の三つの評価実験を行う．

**評価実験 1** 「家庭内危険行動」に対して「危険な理由」は適切であるか

**評価実験 2** 「危険な理由」に対して「根拠文」は適切であるか

**評価実験 3** 専門文書から検索された根拠文は正確であるか

表 3: 評価実験 1 の結果

| 家庭内危険行動 | 平均 | 標準偏差 |
|---|---|---|
| R1 | 4.1 | 1.1 |
| R2 | 3.6 | 1.4 |
| R3 | 3.0 | 1.6 |
| R4 | 3.1 | 1.7 |
| R5 | 3.5 | 1.5 |
| R6 | 3.5 | 1.4 |
| R7 | 3.1 | 1.5 |
| R8 | 4.3 | 1.1 |
| R9 | 4.1 | 1.1 |
| 全体 | 3.6 | 1.4 |

表 4: 評価実験 2 の結果

| 家庭内危険行動 | 平均 | 標準偏差 |
|---|---|---|
| R1 | 2.9 | 1.7 |
| R2 | 2.8 | 1.7 |
| R3 | 1.6 | 1.0 |
| R4 | 2.5 | 1.6 |
| R5 | 2.5 | 1.5 |
| R6 | 2.6 | 1.5 |
| R7 | 3.2 | 1.6 |
| R8 | 3.5 | 1.4 |
| R9 | 2.1 | 1.3 |
| 全体 | 2.6 | 1.6 |

評価実験 1 と評価実験 2 に関しては，被験者アンケートを行う．被験者は大学生 15 人である．被験者の認知バイアスによる影響を抑えるために，被験者には使用した LLM を伝えず，質疑応答システムの返答に関して評価するように伝える．また，被験者の主観により評価が揺れないように，評価の基準を被験者に提示した．評価実験 1 の基準は,「“指定した家庭内危険行動について，日常生活で想定される危険な理由” を適切とする」とする．評価実験 2 の基準は,「“文の種類が異なっていても，根拠文から行動が危険である理由を推測できる文” を適切とする」とする．

評価指標には，リッカート尺度に基づく 5 段階評価を使用する．具体的には，適切，やや適切，どちらとも言えない，やや不適切，不適切のうち，最も当てはまる評価値を被験者に回答してもらう．評価実験 1 では九つの家庭内危険行動に対して危険な理由が適切であるかの評価値を回答してもらう．評価実験 2 では四つの専門文書それぞれについて，危険な理由に対する根拠文が適切であるかの評価値を回答してもらう．評価実験 1 と評価実験 2 では，適切から不適切の評価値を 5 から 1 の数値に置き換え，それぞれの家庭内危険行動ごとに被験者による評価の平均値と標準偏差を小数第一位まで求める．評価実験 3 に関しては，客観的に判断できる基準を決めて筆者が評価する．以下に基準となる条件を示す．

**条件 1** 専門文書中の記述を一言一句変わらず，一文で出力できている

**条件 2** 高齢者の転倒・転落に関する家庭内危険行動の理由の根拠を出力できている

**条件 3** 性別の指定や病気・怪我を患っている人など，一部の高齢者に関する情報ではない

**条件 4** Elasticsearch の結果が LangChain の結果のメタデータとして正しい

条件 1 に関して，箇条書きの行頭文字や段落番号は，欠損していても条件を満たしていると定義する．条件 4 に関して，メタデータであるページ数は，文書中に記載されたページ数ではなく，PDF ファイルが示すページ数と定義する．そのため，メタデータのページ数と文書中の実際のページ数が異なる場合がある．

### 4.4 評価実験結果

評価実験 1 の結果を表 3 に示す．評価実験 1 の全体の平均値は「3.6」となり，“どちらとも言えない” と “やや適切” の中間に位置する結果となった．

評価実験 2 の結果を表 4 に示す．評価実験 2 の全体の平均値は「2.6」となり，“やや不適切” と “どちらとも言えない” の

表 5: 評価実験 2 の専門文書別の結果

| 家庭内危険行動 | 専門文書 1 | | 専門文書 2 | | 専門文書 3 | | 専門文書 4 | |
|---|---|---|---|---|---|---|---|---|
| | 平均 | 標準偏差 | 平均 | 標準偏差 | 平均 | 標準偏差 | 平均 | 標準偏差 |
| R1 | 3.3 | 1.6 | 2.3 | 1.6 | 2.1 | 1.6 | 3.8 | 1.3 |
| R2 | 3.5 | 1.6 | 1.6 | 1.1 | 3.0 | 1.9 | 2.9 | 1.7 |
| R3 | 1.7 | 1.1 | 1.8 | 1.2 | 1.2 | 0.6 | 1.9 | 1.1 |
| R4 | 3.5 | 1.6 | 1.3 | 0.8 | 1.9 | 1.3 | 3.4 | 1.5 |
| R5 | 3.5 | 1.4 | 2.2 | 1.3 | 2.2 | 1.6 | 2.2 | 1.6 |
| R6 | 2.5 | 1.4 | 1.7 | 1.2 | 3.9 | 1.1 | 2.3 | 1.5 |
| R7 | 3.5 | 1.6 | 1.5 | 0.8 | 4.1 | 1.3 | 3.5 | 1.6 |
| R8 | 3.9 | 1.2 | 3.1 | 1.6 | 3.7 | 1.5 | 3.5 | 1.5 |
| R9 | 1.6 | 1.1 | 1.8 | 1.0 | 2.9 | 1.4 | 2.2 | 1.5 |
| 全体 | 3.0 | 1.6 | 1.9 | 1.3 | 2.8 | 1.7 | 2.9 | 1.6 |

表 6: 評価実験 3 の結果

| 家庭内危険行動 | 専門文書 1 | | | | 専門文書 2 | | | | 専門文書 3 | | | | 専門文書 4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | C1 | C2 | C3 | C4 | C1 | C2 | C3 | C4 | C1 | C2 | C3 | C4 | C1 | C2 | C3 | C4 |
| R1 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| R2 | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| R3 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ |
| R4 | | ✓ | | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| R5 | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| R6 | | ✓ | | ✓ | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| R7 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| R8 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| R9 | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ |

中間に位置する結果となった．また，評価実験 2 の専門文書別の結果を表 5 に示す．表 5 より，四つの専門文書のうち最もスコアが高い根拠文を，危険な理由の適切な根拠文として選定すると，スコアは 2.9〜4.1 の間に収まる．

評価実験 3 の結果を表 6 に示す．表 6 では，条件 1 から条件 4 を C1 から C4 で表記する．条件を満たす項目にそれぞれチェックマークを付けた．四つすべての条件を満たす根拠文を正確と定義する．正確な根拠文は，24/36（66.7%）となった．条件 2 と条件 4 に関しては，すべての根拠文において条件を満たす結果になった．

### 4.5 考察

評価実験 1 の結果より，家庭内危険行動の理由を一言で表現することは難しく，被験者が適切と認識できるには限界があるという考察が得られた．

評価実験 2 の結果より，提案システムが提示した根拠文は，家庭内危険行動が危険であることの信頼性を示す，一般的な根拠文であり，根拠として扱うには具体性に欠ける内容も多いという考察が得られた．一般的な根拠文には，足元のトラブルやバランス能力の低下などによる転倒の危険性が記述されていた．また，専門文書によって精度が変わるため，複数の専門文書で根拠付けを行い，適切さを人間が行うのであれば，文章生成 AI の出力に根拠文を付与することも可能であると考えられる．

評価実験 3 の結果より，専門文書中の記述を一言一句正確に出力できないものの，欠損箇所は接続詞などが多く，意味的には問題ない文が多かった．また，専門文書中の記述を一言一句変わらず出力することは難しいという考察が得られた．

## 5. おわりに

本研究では，文章生成 AI が生成した，高齢者の家庭内危険行動の理由に類似する文を，専門文書から RAG と Elasticsearch を用いて検索し，根拠文としてユーザに提示するシステムを提案した．

本研究の今後の課題として，様々な形式で記述された専門文書や知識グラフを根拠の知識源とすることや，他の LLM との比較が挙げられる．

## 謝辞

本研究成果の一部は，国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の委託業務 (JPNP20006) の結果得られたものです．本研究は JSPS 科研費 23K11221 の助成を受けたものです．

## 参考文献

[Egami 23] Egami, S., et al.: Synthesizing Event-Centric Knowledge Graphs of Daily Activities Using Virtual Space, *IEEE Access*, Vol. 11, pp. 23857–23873 (2023)

[Lewis 21] Lewis, P., et al.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, ArXiv, abs/2005.11401 (2021)

[鵜飼 22] 鵜飼 孝典 他：高齢者の家庭内事故予防に役立つ AI システムの開発 ―産業版ナレッジグラフ推論チャレンジに向けて―, 人工知能学会第二種研究会資料，SIG-SWO-056-15 (2022)

[江上 22] 江上 周作 他：家庭内の事故予防に向けた合成ナレッジグラフの構築と推論, 人工知能学会第二種研究会資料, SIG-SWO-056-14 (2022)

[浅野 23] 浅野 歴 他：知識グラフと GPT を用いた家庭内の危険行動の検知と説明, 2023 年度人工知能学会全国大会（第 37 回），3G1-OS-24a-03 (2023)