子どもの事故事例に基づく RAG 型警告システムの開発

平居珠実[†1]　岡夏樹[†2]　田中一晶[†1]

概要：保護者が子育てをする上で，子どもの事故を防止するための知識が必要となる．このような知識は専門家からアドバイスを得る等して身につけることができるが，座学で得られる知識には限りがある．そこで，実際の場面から注意すべきことを学べる方法として，子どもを撮影した写真に対して警告文を生成する警告システムを開発した．被験者アンケートを用いた評価実験の結果から，システムの構築に事前学習済み大規模言語モデルを用いることで，多くの場合，被験者より適切な警告文生成ができることがわかった．さらに，入力写真に類似する状況で過去に起こった事故を検索し，その事故の情報を写真と共に与える Retrieval Augmented Generation を行うことで，画像の状況をより正確に反映した警告文生成が可能になることもわかった．

1. はじめに

子育てにおける不安や悩みは深刻な問題である．これらが虐待に発展する事例も多く，こども家庭庁の調査では児童虐待等の問題も 30 年連続で増加しており，2022 年度には，児童相談所での対応件数が過去最高である約 22 万件にも上っている．特に死亡事例のうち約 70%は養育者の育児不安もしくは養育の力の低さからくる心理的・精神的問題が原因であることが判明している．このような育児不安等の原因の 1 つとして，子育てにおいては知識不足が子どもの不慮の事故や体調不良に直結することが挙げられる．親の知識不足を改善するためには，保健委員会の職員による各家庭の安全点検や，小児科医によるカウンセリングなどを含む傷害予防プログラムが効果的である[1].

しかしながら，専門家が各家庭にアドバイスをするのは人手がかかる．そこで，専門家の代わりとなるシステムが開発されており，子どものよじ登りをシミュレーションする研究等が行われている[2]．しかしながら，子どもに起こり得る事故は転落以外にも誤飲や火傷，窒息など種類が多く，転落事故のみの予測では子どもの事故を防止する知識として不十分である．様々な種類の事故を 1 つのシステムで予測できればより効率的に親が子育てにおいて注意すべきポイントを学べると考えた．本研究では，大規模言語モデルと事故データベースを組み合わせる事によって，子どもの状況を撮影した写真から子どもに起こりそうな事故の種類を予測し，警告文の生成を行う警告システムを開発した．

2. 関連研究

子育てを支援するシステムの先行研究として，スマートスピーカーを用いて家族間のコミュニケーションを促進したり，親が立てた目標やすべき行動について音声でサポートしたりするシステムが開発されている[3]．このシステムを用いることで，スマートスピーカーのアシストにより，よりスムーズな子育てが可能となるが，あくまでも行動したり目標を設定したりするのは家族のメンバーであり，負担の軽減や知識の提供等のサポートは行うことができない．

子育てに従事する保護者の負担を軽減するシステムの研究として，自律的に遊び相手となって子どもの注意を引くものが挙げられる[4][5]．しかしながら，ロボットが遊び相手として機能している間は親の負担が軽減されるかもしれないが，その効果は飽きによって次第に薄れていき，長時間使い続けることはできない．そのため、可能な限り長時間のインタラクションが生じるように話す・遊ぶ手法が研究されている[5]．また，遠隔操作によって子どもと遊んだり見守ったりするロボットも提案されている[6][7]．しかしながら，ロボットや遠隔地の協力者に代わりに子守を任せるのはあくまでも一時的な手段であり，親が子どもの安全のための知識を身につける必要性は変わらない．

子どもがいる環境内の危険要因を知らせるシステムとして，子どものよじ登りをシミュレーションする研究が行われている[2]．このシステムを使用すれば，環境に潜む危険要因を排除することができ，根本的な事故防止に繋がる．しかしながら，気を付けるべき危険は多岐にわたり，1 種類の事故についてしかシミュレーションできないのは不便である．

このような背景に加え，共働きの親が急増したことでオンラインの子育て資料の需要が高まっており，これに伴い，近年子どもをもつ母親の間ではインターネットの利用が増加している[8][9]．子育ての情報源としてインターネット上の web サイトやブログ，SNS を使用する幼い乳児の親が多いことが先行研究の調査でわかっている[10][11]．しかしながら，インターネット上の情報には信頼性の無い情報が混じっていたり，価値のある情報かどうか不明であったりと，問題点も多い．

†1 京都工芸繊維大学
†2 宮崎産業経営大学

そこで，子育てツールの 1 つとしてアプリを積極的に使用する親も多く[12][13][14]，アプリの信憑性についての調査も行われている[15]. 実際に，アプリの利用が母親の役に立つという調査結果も存在する[16]．保護者向けに提供されているアプリとして，授乳時間や排泄の記録を付けたり，身長や体重の推移をわかりやすく纏めるトラッキングアプリが複数提供されており，子どもの健康管理に役立てられている．他にも情報共有アプリや睡眠補助アプリ，写真共有アプリなども提供されている．保護者が能動的に必要な情報の記録や収集をするために用いるアプリの他に，最近では Baby Buddy[17]や Rational Parenting Coach (RPC)[18]等，アプリがユーザを関連コンテンツに誘導するものも開発されている．

本研究では，このような保護者に必要な知識を積極的に提供するシステムとして，子どもの写真に対して警告文を生成するシステムを構築することで，より効率的な知識の獲得が可能になると考えた．保護者が家庭やよく遊ぶ場所で撮影した自身の子どもの写真を本システムに入力すると，その場面でどのようなことに気を付けるべきであったかというフィードバックが出力され，実践的に子どもの事故防止に必要な知識を学ぶ事ができる．既存のシステムと異なり，保護者に情報を提供する手段として，既製コンテンツに誘導するのではなく，状況や環境に即した警告文を生成するため，各ユーザが必要な情報を効率的に集められる．危険予測システムの構築方法については次章で示す．

3. RAG 型危険予測システムの構築

画像と言語を入力として推論を行う大規模言語モデルとして，ChatGPT[19]や Gemini[20]，LLaVA[21]などが挙げられる．しかし，これらの大規模言語モデルは一般的な画像とそれに関する会話を学習データとしており，知識集約的な専門性のあるタスクには適さない．また，事故に遭う可能性のある子どもを撮影した写真に対する警告というデータは入手しづらく，ファインチューニングによって子どもの事故に関する知識を付け加えるのは困難である. そこで，本研究では，入力内容に関連する外部データを追加で与えることで言語モデルの推論精度を高める Retrieval Augmented Generation (RAG)[22]を用い，事前学習済みモデルを知識集約的なタスクに対しても適切な推論を行えるようにすることで危険予測システムを構築する．

本研究では，大規模言語モデルとして画像と言語を扱う事前学習済みモデルである LLaVA を使用した. 事前学習済み大規模言語モデルとしては ChatGPT のようなクラウド上のモデルが使用されることが多い．しかしながら，オンライン環境で動かされるモデルは入力データを学習データベースとして蓄積する等の情報漏洩のリスクがあるため，過去の子どもの事故に関するデータや子どもが写り込んだ写真のようなプライベートな情報を用いた推論を行うのには適さない．一方，LLaVA はモデルをローカル環境にダウンロードした上で使用できるため，入力内容がシステム外に漏洩するリスクが低く，秘匿性が求められる情報を扱うのには適している．そのため，警告システムの構築には LLaVA を使用した．

プロンプトに付け加える外部データとして国立成育医療研究センターが過去 6 年にわたって 22,011 件の事故の情報を収集した事故データベースを使用した．このデータベースには過去に発生した事故について，事故に遭った子どもの年齢や性別，事故の発生場所，原因となった物などの情報がテキストで格納されている．事故データベースに記載されている事例の一例を表 1 に示す.

データベースの中から画像と類似した状況で過去に起こった事故事例を検索するために，画像と言語の共通空間モデルである CLIP[23]を用いたシステムを構築した．あらかじめ，事故データベース内に記載されている事故が起きた場所や直前の行動等の情報を用いて，各事故の発生直前の状況を表す文を生成した．写真が入力されると，写真と各状況文の類似度スコアを CLIP で計算し，類似度が上位 5 件の事例の情報を LLaVA に与えるようにした．情報は <similar accidents> </similar accidents>タグで囲み，プロンプトの文末に付け加えた．このように，子どもの画像と警告文生成を指示するプロンプトを入力として与えると，CLIP を用いて画像に類似した状況で過去に起こった事故を検索し，画像と過去の事故の情報を用いて，起こりそうな事故の予測及び警告文の生成を行う RAG 型システムを開発した．

4. 実験

4.1 仮説

大規模言語モデルは膨大なデータを学習しているため，人間よりも知識量が多い．そのため，子どもの危険を警告するタスクにおいて，言語モデルを使用したシステムは人間よりも適切な警告文を生成できると考えた．さらに，画像内の子どもと類似した状況で起こった事故の情報を言語情報として与えることによって，より適切な警告内容が得られると考えた．したがって，本研究の仮説は以下の 2 つである.

仮説 1：言語モデルを使用したシステムが生成する警告文は人間が作成する警告文よりも適切である

仮説 2：画像のみを入力した場合と比較して，画像から検索した類似状況下での事故情報を加えた場合，警告文の適切性は高くなる

4.2 実験方法

実験に使用するモデルとして以下の 2 種類のモデルを構築した.

ベースラインシステム : LLaVA が画像から警告文を生成するシステム

表 1　事故データベース内の事故事例の一例

項目	記載内容
性別	女
発達段階	よちよち歩きができる
事故発生月	11
月齢	13
年齢	1 歳 1 ヶ月
事故発生時間	22
事故の種類	転倒
事故の原因になったもの	木製のローテーブル
傷害の原因になったもの	木製のローテーブル
事故が起きた場所	家庭
家庭	自宅
家庭の詳細	居間
直前の行動	歩いていた
怪我の種類	挫創
治療状況	経過観察(治療不要)
事故の詳細	母は不在だった．23 : 00 帰宅して歯磨きをしていて口腔内の出血，創に気づいた．留守番をしていた父を問いただしたところ，転倒，打撲を知ることになった．
怪我の部位	口・口腔・歯

RAG 型システム： CLIP で検索した事故事例を LLaVA への入力に加えたシステム

両システムの構築において LLaVA への指示(インストラクション，ロール，プロンプト）は子どもの事故予測に適したものを予め設定しておき，固定とした．指示は子どもの様々な場面を写した開発用の 12 枚の写真を用いて最も事故予測に適するものを検討して作成した．与えた指示は表 2 に示す通りである．

実験に用いる画像として，子どもの日常の様々な場面を映した写真を屋内の写真 7 枚，屋外の写真 7 枚の計 14 枚用意し，ベースラインシステムと RAG 型システムの 2 つのシステムに各写真に対する警告文を生成させた．これらの写真は，似た状況で発生した事故が少なくとも 1 種類以上データベース内に含まれている場面の写真である. なお, LLaVA の出力にはランダム性があるが，公平性を期すため, 全ての写真において 1 回のみ警告文を出力させ，それを実験に使用する警告文とした．なお，警告文は英語で出力されるため，実験ではこれらを日本語に自動翻訳したものを使用した．

評価は被験者アンケートで行った．被験者が作成した警告文と 2 つのシステムが生成した警告文の計 3 種類に対して，適切性を評価させた．人間と両システム間，ベースラインシステムと RAG 型間で適切性に差があるかを調査した．

4.3 実験手順

アンケートは 2 パートに分けて行い, パート 1 では 14 枚の写真を被験者に見せ，各写真に対して警告文を作成させた．警告文の評価の際に被験者本人が作成した警告文を客観的に評価できるようにするため，パート 1 の実施から 2 日以上時間をあけ，被験者がどのような警告文を作成したか記憶が薄れたタイミングで，パート 2 を行った．パート 2 では，写真とそれに対してパート 1 で自身が作った警告文，ベースラインシステムが生成した警告文，RAG 型システムが生成した警告文の 3 種類の警告文を被験者に見せた. バイアスがかからないよう, ベースラインと RAG 型の 2 つのシステムが生成した警告文を被験者が区別できないように提示する順番をランダムに入れ替えた．本研究のシステムは保護者に写真の状況で起こりそうな事故を教え，防止策を学習させることを目的としているため，適切に危険を警告できているか（適切性）について評価させた．この質問項目について，「1:全くあてはまらない」～「7:非常によくあてはまる」の 7 段階で評価させた.

表 2　LLaVA への指示

インストラクション	A chat between a parent and an experienced childcare worker. The childcare worker has seen many children's accidents. The childcare worker has extensive knowledge of children's accidents. The mission of the childcare worker is to keep children safe. The childcare worker uses the knowledge to give accurate and concise advice to questions of the parent.
ロール	Parent, Childcare worker
プロンプト	"Look at every inch of what is in this picture. Determine which objects in the photograph are most likely to be a factor in the accident, predict an accident that might happen to the child and make a warning statement that should be communicated to parents to prevent that accident in <warning></warning> XML tags.

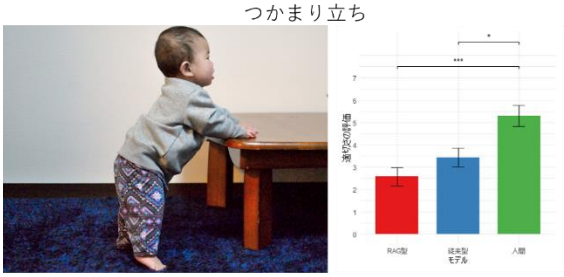
被験者として著者の所属する京都工芸繊維大学の学部生及び修士課程の 20 代の学生を男女 7 名ずつ，計 14 名集めた．14 名の被験者はいずれも育児経験がなく，本システムの利用者として想定される子どもの事故防止について学習中で知識の乏しい保護者を対象とした時と近い評価結果が得られると考えられる. 14 枚の写真を評価する慣れの効果を除外するため，被験者ごとに異なる順番で写真を提示した.

5. 実験結果

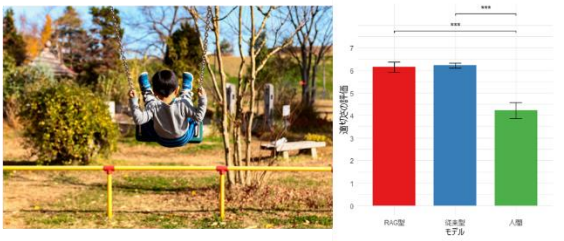
14 枚の写真に対して, 人間, ベースラインシステム, RAG 型システムそれぞれが作った警告文の適切性の評価は図 1 の通りになった．適切性について分散分析し，主効果が認められた場合には TukeyHSD 法で多重比較した.

その結果，屋内の写真に関しては，つかまり立ちの写真で，評価に有意差が見られ（$F(2, 39) = 10.13, p < .001$），事後検定により，人間とベースライン（$p<.05$），人間と RAG 型（$p<.001$）に有意差が見られ，いずれも人間の方が評価が高かった．食事の写真でも有意差が見られ（$F(2, 39) =3.974, p < .05$），人間と RAG 型の間に有意差が見られた（$p<.05$）．一方で，キッチンの写真では（$F(2, 39) = 6.722,$

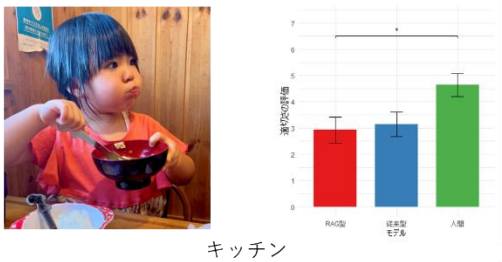
つかまり立ち



ブランコ

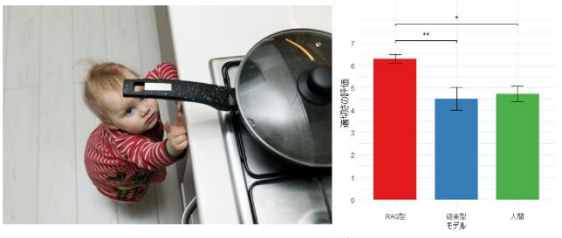


食事



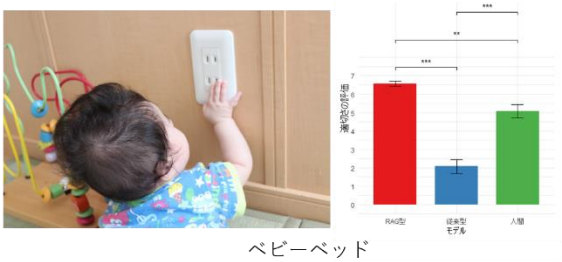
木登り

キッチン

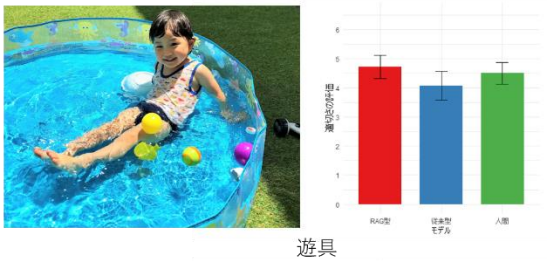


かけっこ

コンセント

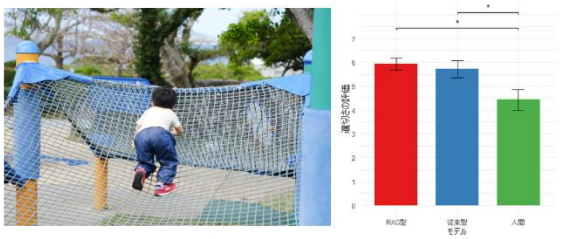


プール

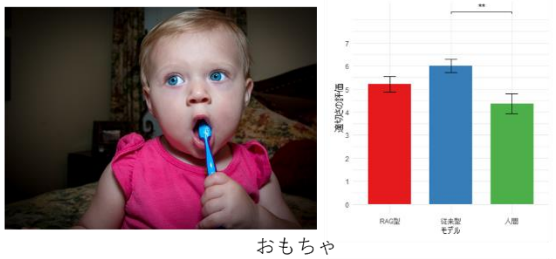


ベビーベッド

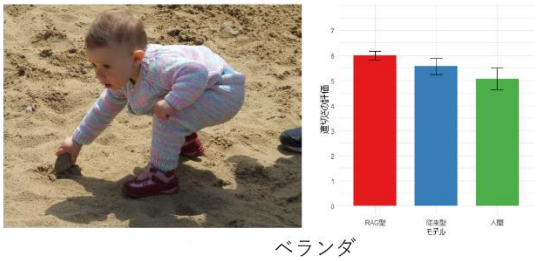
遊具



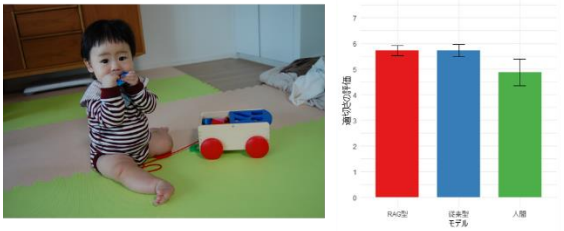
歯磨き



砂遊び



おもちゃ



ベランダ

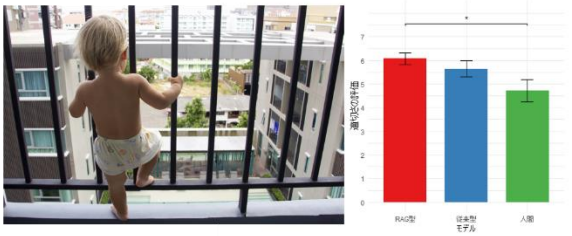


図 1　各写真の評価結果

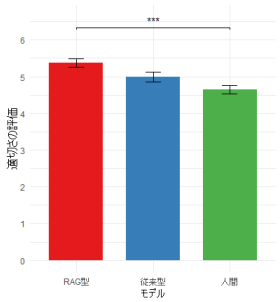


図 2　写真全体の評価結果

p< .001），人間と RAG 型（p <.05），ベースラインと RAG 型（p <.001）に有意差が見られ，RAG 型が優れていた．コンセントの写真は（F(2, 39) = 55.89, p < .001），人間とベースライン（p <.001），人間と RAG 型（p <.01），ベースラインと RAG 型 (p <.001) の全ての間に有意差が見られ，RAG 型が最も優れていた．ベビーベッドでは（F(2, 39) = 4.74, p < .05），人間とベースラインの間に有意差が見られ（p <.05），ベースラインが優れていた．歯磨きの写真でも（F(2, 39) = 5.151, p < .05），人間とベースラインの間に有意差が見られ（p <.01），ベースラインが優れていた．おもちゃの写真では（F(2, 39) = 1.983, p = .151）有意差は見られなかった．

屋外の写真に関しては，ブランコの写真は（F(2, 39) = 20.42, p < .001）では，人間とベースライン（p <.001），人間と RAG 型の間に有意差が見られ（p <.001），どちらもシステムが高評価であった．木登りの写真では（F(2, 39) = 4.305, p < .05），人間と RAG 型の間に有意差があり，RAG が優れていた（p <.05）．かけっこの写真では（F(2, 39) = 14.1, p < .001），人間とベースライン（p <.001），人間と RAG 型（p <.001）に有意差が見られ，いずれもシステムが優れていた．遊具の写真でも（F(2, 39) = 5.346, p < .01），人間とベースライン（p <.05），人間と RAG 型（p <.05）に有意差が見られ，いずれもシステムが優れていた．ベランダの写真では（F(2, 39) = 3.701, p < .05），人間と RAG 型の間に有意差が見られ RAG 型が優れていた（p <.05）．プールの写真（F(2, 39) = 0.589, p = .56）と砂遊びの写真（F(2, 39) = 1.954, p = .155）について有意差は確認されなかった．

全画像の評価を纏めて分散分析した結果は図 2 の通りである．その結果，有意差が見られ（F(2, 585) = 9.256, p < .001），事後検定により人間と RAG 型の間に有意差が見られ（p <.001），RAG 型が高評価であったことがわかった．

6. 考察

6.1 人間とシステムの警告文の適切性の比較

適切性の評価では，多くの写真についてベースライン，RAG 型の方が人間よりも適切な警告文を生成できていることが分かった．ベースラインと RAG 型両方の警告文が人間のものよりも評価が低かったつかまり立ちの写真について両 AI システムが適切な警告文を生成できなかった原因として，今回の実験で使用した CLIP 及び LLaVA が主に英語圏の画像と文を学習データとしているため，日本でよく見られる床に座って使うような高さの低い机の写真が学習データに含まれておらず，正確な状況理解ができなかった可能性が考えられる．また，RAG 型の方が人間よりも評価が低かった食事の写真やコンセントの写真ではハルシネーションが起こっており，写真内のフォークをナイフと誤認識したり，コンセントを電灯スイッチとし誤認識したりしていた．このように，学習データの画像と実験に用いた写真の文化の差や，ハルシネーションが原因で，一部の写真ではシステムは人間よりも劣る警告文しか生成できなかったが，それ以外の写真においては，人間よりも適切な警告文を生成することができた．したがって，言語モデルを使用したシステムが生成する警告文は多くの写真で，人間が作成する警告文よりも適切であるという仮説 1 の通りの結果が得られた．

6.2 事故事例の有無による警告文の適切性の比較

また，2 枚の写真でベースラインよりも RAG 型の警告文が適切だと評価されており，ベースラインの警告文が RAG 型よりも適切だと評価された写真は存在しなかった．RAG 型が優れた評価を得た 2 枚の写真では，ベースラインが状況把握において物体を誤認識し，写真の状況にそぐわない警告文を生成していたが，RAG 型では検索によって得られた過去の事例として写真の状況に関連するものを選べていたため，誤認識が起こらず，状況に見合った警告文を生成できていた．このように，RAG 型にすることで状況把握の失敗を修正することができており，画像のみを入力した場合と比較して，画像から検索した類似状況下での事故情報を加えた場合，警告文の適切性は高くなるという仮説 2 の通りの結果が得られた．

6.3 今後の展望

今回の実験により，子どもの写真に対して危険警告をするというタスクにおいて，既存の大規模言語モデルに過去の事故事例を外部テキストデータとして付け加えることで，人間や大規模言語モデル単体よりも適切な警告文を生成できることがわかった．本システムは複数のモジュールを組み合わせて構築しているため，モジュールを入れ替えることが容易であり，今後技術の向上に伴い各モジュールをより高性能なものに入れ替えることで，本システムの性能を向上させることができる．また，事故事例は外部データとして扱われており，追加や編集が容易である．今後の展望として，生活環境の変化による新たな種類の事故に対応できるように，最新の事故事例を収集することが挙げられる．今回使用した事故データベースは医療機関等の協力のもと収集されたデータでるため，

今後も医療機関等で事故に遭った子どもの詳細情報を記録することで，本システムの構築に必要なデータの収集が可能である．

7. まとめ

本研究では，子どもを撮影した写真から事故防止のために周囲の大人が気を付けるべきことを学べる RAG 型警告システムを構築し，言語モデルに過去の事故事例を与える事で出力がどのように変化するかを被験者実験により調査した．その結果，言語モデルを使用したシステムは人間よりも適切に警告文を生成でき，事故事例を参照することで誤認識の少ないより適切な警告文が生成できることがわかった．本研究で構築した子どもの事故事例に基づく RAG 型危険予測システムは子どもの写真に対して適切な警告文を生成でき，子育てにおいて気を付けるべきことを人間に教えるという目的を果たし得る精度であるといえる．

謝辞 本研究のシステム構築に用いた事故データベースを提供してくださった国立成育医療研究センター，LLaVA の場面設定を作成するにあたって言語学の観点から助言をくださった京都工芸繊維大学の深田智教授に，謹んで感謝の意を表する．本研究は，JSPS 科研費 JP22K12126，JP24K00066 の支援を受けた．

参考文献

[1] B Guyer, S S Gallagher, B H Chang, C V Azzara, L A Cupples, and T Colton, "Prevention of Childhood Injuries: Evaluation of the Statewide Childhood Injury Prevention Program (SCIPP)." American Journal of Public Health, vol.79, no.11, pp.1521-1527, 1989.
[2] Tsubasa Nose, Koji Kitamura1, Mikiko Oono, Yoshifumi Nishida, Michiko Ohkura, "Data-driven child behavior prediction system based on posture database for fall accident prevention in a daily living space," Journal of Ambient Intelligence and Humanized Computing, vol.11, pp.5845-5855, 2020.
[3] E. Beneteau, A. Boone, Y. Wu, J.A. Kientz, J. Yip, A. Hiniker, “Parenting with Alexa: Exploring the Introduction of Smart Speakers on Family Dynamics,” Proceedings of the 2020 CHI conference on human factors in computing systems, pp.1-13, 2020
[4] 二又川 求哉ら，子育て支援ロボットの開発，ロボティクス・メカトロニクス講演会講演概要集, 2P2-B16, 2018
[5] Altion Simo et.al., A Humanoid Robot to Prevent Children Accidents, Interactive Technologies and Sociotechnical Systems, 476–485, 2006
[6] K Abe, M Shiomi, Y Pei, T Zhang, N Ikeda, T Nagai, "ChiCaRo: tele-presence robot for interacting with babies and toddlers," Advanced Robotics, vol.32, no.4, pp.176-190, 2018.
[7] 遠隔育児支援ロボット ChiCaRo，電気通信大学/大阪大学長井研究室. http://www.rlg.sys.es.osaka-u.ac.jp/chicaro/（参照 2024-10-28）
[8] S. Shorey, Y.P.M. Ng, D.B. Danbjørg, C.L. Dennis, E. Morelius, “Effectiveness of the ‘Home-but not Alone’ mobile health application educational programme on parental outcomes: a randomized controlled trial,” Journal of advanced nursing, vol.73, no.1, pp.253-264, 2016.
[9] A. Rathbone, J. Prescott, “I Feel like A Neurotic Mother at Times”-a mixed methods study exploring online health information seeking behaviour in new parents”, mHealth, vol.5, 2019.
[10] R.Y. Moon, A. Mathews, R. Oden, R. Carlin, “Mothers’ Perceptions of the Internet and Social Media as Sources of Parenting and Health Information: Qualitative Study”, Journal of Medical Internet Research, vol.21, no.7, 2019
[11] D. Lupton, S. Pedersen, G.M. Thomas, “Parenting and Digital Media: From the Early Web to Contemporary Digital Society,” Sociology compass, vol.10, no.8, pp.730–743, 2016.
[12] D. Lupton, “The use and value of digital media for information about pregnancy and early motherhood: a focus group study,” BMC Pregnancy Childbirth, vol.1, no.1, 171.doi: 10.1186/s12884-016-0971-3, 2016.
[13] Pehora C, Gajaria N, Stoute M, et al. “Are Parents Getting it Right? A Survey of Parents' Internet Use for Children's Health Care Information,” Interactive Journal of Medical Research, vol.4, no.2, 2015.
[14] Chin K. Prospective data collection for feeding difficulties and nutrition [dissertation]. Boston, MA: Boston University, vol.16, no.1, 2018.
[15] A. Virani, L. Duffett-Leger, N. Letourneau, “Parenting apps review: in search of good quality apps,” mHealth, vol.5 ,2019.
[16] Zhao J, Freeman B, Li M. "How do infant feeding apps in China measure up? A content quality assessment," JMIR mHealth and uHealth, vol.5, vol.12, e8764, 2017.
[17] A. Rhodes, S. Kheireddine, A.D. Smith, “Experiences, Attitudes, and Needs of Users of a Pregnancy and Parenting App (Baby Buddy) During the COVID-19 Pandemic: Mixed Methods sStudy,” JMIR mHealth and uHealth, vol.8, no.12, e23157, 2020.
[18] O.A. David, “The rational parenting coach app: rethink parenting! a mobile parenting program for offering evidence-based personalized support in the prevention of child externalizing and internalizing disorders,” Journal of Evidence-Based Psychotherapies, vol.19, no.1, pp.97-108, 2019.
[19] Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al, “Gpt-4 technical report,” arXiv preprint arXiv:2303.08774, 2023.
[20] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al, “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,” arXiv preprint arXiv:2403.05530, 2024.
[21] H. Liu, C. Li, Q. Wu, and Y. J. Lee, " Visual instruction tuning, " arXiv preprint arXiv:2304.08485, 2023.
[22] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. K¨uttler, M. Lewis, W.-t. Yih, T. Rockt¨aschel et al., “Retrieval-augmented generation for knowledge-intensive nlp tasks,” Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.
[23] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, "Learning Transferable Visual Models From Natural Language Super", Proceedings of the 38th International Conference on Machine Learning, pp.8748-8763, PMLR 139:8748-8763, 2021.