

# マルチモーダル大規模言語モデルを用いた子どもの危険警告システムの設計指針の導出

平居珠実<sup>†1</sup> 岡夏樹<sup>†2</sup> 田中一品<sup>†3</sup>

**概要**：子どもの事故を回避するために開発されているシステムは、特定の事故だけを対象とするか、一般的な環境での対策を提示するものがほとんどである。これに対し、本研究では、マルチモーダル大規模言語モデル（LMM）を活用することで、個々の子どもの育児環境で起こり得るより多様な事故を予測・警告することができると考え、次の2種類のプロトタイプシステムを構築した。システム A：入力した育児環境の写真とプロンプトに基づいて、起こり得る事故と回避方法を説明する警告文を出力する。システム B：入力写真の状況から子どもの事故データベースを検索し、プロンプトと共に事故事例も入力する。育児の未経験者と専門家を対象とした評価実験を通して、事故事例を検索・入力して得られた警告文によって一般的な事故を警告し、さらに事故事例を入力せずに得られたその環境特有の稀な事故の警告を提示することが有効である等の設計指針を明らかにした。

## 1. はじめに

不慮の事故で子どもが怪我をしたり命を落としたりする事例は絶えない。子どもの事故の種類は、ボタン電池の誤飲やベランダからの転落、ヘアアイロンによる火傷など、多岐に渡り[1]、中には子どもが命を落とした事例も報告されている。保護者は子どもの事故防止に努めるが、あらゆる種類の事故に対して対策をとるのは難しい。そのため、子どもの事故防止をサポートするシステムの開発は大きな課題といえる。

保護者が様々な事故に対策できるように手助けする手段の1つとして、保護者に知識を提供する手法が挙げられる。保護者に事故防止のために必要な情報を提供し知識を身につけさせることで、保護者が様々な事故を防ぐための対策をとれるようにする。このような手法の一例として、発育段階に応じてコラムを提供するアプリや[2]、よく起こる事故についての学習用のアニメーション動画を提供する Web サイト[3]等が挙げられる。これらのシステムはコラムや動画、統計結果などを保護者に提供することで知識を向上させる。一方で、これらのシステムで提供されるコンテンツは他人の家庭や、普段利用している施設・公園ではない場所の事故をもとにしているため、対策が保護者の環境にも合致するとは限らない。

本研究では、保護者の子育て環境に対してアドバイスを提供するシステムを構築すれば、保護者が効率的に必要な情報を収集できると考えた。数値や統計ではなく、専門知識のない保護者にもわかりやすいような文で情報を提供するために大規模言語モデルが使用できると考えた。さらに、言語に加えて画像も扱えるマルチモーダルモデルなモデルを使用することで、子どもが置かれている環

境を写真で入力することができ、より扱いやすいシステムが構築できる。しかし子どもの事故についてのデータは一般的な会話データと異なり収集が難しく、子どもの事故について学習したマルチモーダルモデルを作成するのは困難である。そこで本研究では、一般的なデータセットから構築された学習済み大規模マルチモーダルモデル（LMM）に外部データとして子どもの事故に関する情報を付け加えて専門性を高めることで、入力画像に対して子どもに起こりそうな事故を予測・警告するシステムが構築できると考えた。外部データとしては、過去に国立成育医療センターが収集した事故データベースが使用できそうだと考えた。

本研究では、使用者の育児環境に応じて起こり得る事故を予測・警告するシステムを構築するための設計指針を明らかにするため、以下の2種類のプロトタイプを作成した。システム A：入力した育児環境の写真とプロンプトに基づいて、起こり得る事故と回避方法を説明する警告文を出力する。システム B：入力写真の状況から子どもの事故データベースを検索し、プロンプトと共に事故事例も入力する。育児の未経験者と専門家にこれらのプロトタイプが生成した警告文を評価させる被験者実験を行い、適切性や知識の補完性が高い警告文が生成される条件を調べた。

次節では子どもの事故防止に対する既存のアプローチを紹介し、3節では本研究の提案手法である RAG 型危険警告システムの構造について説明する。また、4 節では RAG 型危険警告システムの評価実験の手順について説明し、5 節で実験結果について示し、6 節で結果についての考察を行う。最後に 7 節で本研究についてまとめる。

†1 京都工芸繊維大学  
†2 宮崎産業経営大学  
†3 甲南大学

## 2. 関連研究

子どもの事故を削減することを目的とした先行研究や既存のシステムで用いられている手法は大きく分けて、遊び相手型、監視型、知識提供型の3つに分けられる。

**遊び相手型**：子どもにロボットと交流させることで危険につながる行動をとらせないようにする手法である。例えば、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]。子どもがロボットに夢中になっている間は子どもの行動を制限できる一方で、飽きによって効果が薄れてしまい、いつ別の行動をとり始めるか予想できないという課題がある。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、この手法ではロボットを操作する人手が必要であり、常に他人が保護者の代わりに子どもを見守り続けるのは現実的ではない。

**監視型**：固定カメラ等を用いて子どもの様子を常に撮影し、危険を検知した場合に保護者に警告を行う手法である。例えば、姿勢推定モデルを用いてあらかじめ指定した危険エリアへの侵入を検出する見守りシステムや[8]、子どもの泣き声を検知し、子どもの位置情報と周辺環境のリアルタイムデータを保護者に送信するシステム等が挙げられる[9]。また、子どもの寝ている姿を映像で見守りうつ伏せ寝を検知する睡眠姿勢検知システムも開発されている[10]。このように個々の事故を検知するシステムが開発されている一方で、多様な事故を検知するシステムは未だに開発されていない。

**知識提供型**：事故の種類や防止策等についての知識を提供することで、保護者が様々な事故に対して防止策をとれるようにする手法である。事前に対策できるため、子どもが思いがけず危険行動とってしまった場合のリスクを減らすことが可能である。また、コンテンツの追加が容易であり、多くの知識を提供しやすい。例えば、子どもの成長過程に合わせて関連コラムを提供するアプリである baby buddy[2]や、子どもの成長過程毎の事故防止策をビデオ教材として Web 上で公開している乳幼児期の事故防止学習ソフト[3]などが挙げられる。また、日常でよく使われる 10 種類の物体に対して保護者が近づくと、過去に子どもがその物体に対してとった予想外な行動の録画が AR で再生されるシステムも開発されている[11]。しかしながら、これらのシステムで提供されるコンテンツは、他人の子どもの事故事例をもとにコンテンツが作成されるため、必ずしも保護者が置かれている状況に即した情報が得られるとは限らないことが課題である。そこで、子どもの行動についてのシミュレーションモデルを構築し、保護者に自身の子育て環境に近い環境で子どもがとりそ

うな行動を直感的に理解させる研究がなされている。例えば、よじ登りや物体を掴むなどの特定の危険行動についてのデータを収集し、それらの危険行動をとりやすい環境を分析してシミュレーションモデルを構築する研究もされている[12][13]。これらの研究ではシミュレーションモデルを構築することで、未知の環境に対しても行動を予測することが可能であるため、保護者の子育て環境に合わせたシミュレーションができる。しかしながら、子どもに起こりえる多様な事故に対してデータを集めシミュレーションモデルを構築するのは困難である。そこで本研究では、保護者の状況に即した知識を獲得でき、かつ多様な事故にも対応可能なシステムとして、LMM を用いたシステムの構築に着目し、LMM をベースとした事故警告システムの有用性を高めるための条件を明らかにする。

## 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マルチモーダルモデルとして、ChatGPT[14]や Gemini[15]、NVLM1.0[16]、LLaVA[17]などが挙げられる。しかし、これらの大規模マルチモーダルモデルは一般的な画像とそれに関するテキストデータを学習データとしており、専門的な知識を扱うタスクには適さない。このような問題に対して、専門的な知識を用いてモデルをファインチューニングすることで、その分野に関するタスクにおいてモデルの文脈理解や専門知識が向上することがわかっている。この手法は、医療分野[18]や医薬品に関するタスク[19]、金融分野[20][21]、法律分野など[22]、様々な分野で広く用いられている。ファインチューニングによってモデルを特定分野に特化させるには、その分野に関するデータセットが必要である。画像と言語の LMM を子どもの事故防止というタスクに特化させるには、事故に遭う可能性のある子どもを撮影した写真とその画像に関する事故情報などをデータセットとして用意する必要がある。しかしながら、事故に遭いそうな子どものデータセットは倫理的な観点から収集が困難である。また、ファインチューニングはモデルにデータセットを再学習させる手法であるため、専門分野のデータを追加するには学習し直す必要がある。一方で、生活環境は日々変化するため、起こり得る事故の種類も変化する。これに対応するためには適宜事故情報を追加する必要があるが、ファインチューニングの手法を用いてモデルを特化させていた場合、事故情報の追加にコストがかかる。これに対し、追加の学習を行うことなくモデルを特定分野に特化させる手法として、Retrieval Augmented Generation (RAG)[23]が提案されている。この手法では、プロンプトに関連する外部データを検索し、プロンプトと共に与えることで大規模言語モデルの推論精度を高める。RAG では、データをモデルに学習させる必要が無いいため、データの追加や入れ替えが

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聞いたところ、転倒、打撲を知ることになった。
怪我の部位	口・口腔・歯

容易である。さらに、大規模言語モデルに外部データを取り入れる場合、RAGの手法を用いた方がファインチューニングよりも性能が上回る事が報告されており[24]、特に専門性の高いタスクほどその傾向が顕著であることが報告されている[25]。このことから、子どもの事故という専門性の高い情報を扱う本研究においてはRAGの手法の方が適切であると考えられる。そこで、本研究では既存のLMMに対してRAGの手法を用いることで子供の事故という専門性の高い情報を扱うシステムを構築する。

LMMに与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。

マルチモーダルなRAGを実現する手法として、大きく分けて以下の3種類が挙げられる。

**共通空間埋め込み**：複数のモダリティの情報を共通のベクトル空間にエンコードし、情報の検索を行う手法である。画像と言語を扱うRAGについて、両方の情報を共通空間ベクトル化するエンコーダを作成し、エンコードしたベクトルの内積からクエリの最近傍のデータを検索する手法などが研究されている[26]。また、画像-言語の既存の学習済み共通空間モデルであるCLIP[27]や、UniBind[28]、OpenFlamingo [29]等を使用して検索する方法も挙げられる。

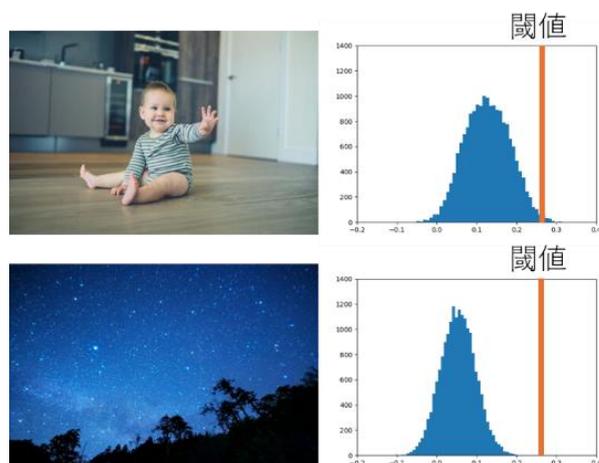


図 1 画像との類似度スコアの分布

**主要モダリティ埋め込み**：複数のモダリティの情報を単一のモダリティの情報に変換して検索を行う手法である。画像のキャプションを生成しテキストとして扱う手法や、入力画像をテキストに変換してドキュメントとの文字列の一致を調べる検索ツールを構築する手法などが研究されている[30]。また、言語情報をテキスト情報に変換する既存の学習済みモデルであるBLIP-2[31]を使用する方法が挙げられる。この手法では画像をテキストにするための前処理が必要となる。

**ベクトルストア**：異なるモダリティ毎にベクトルストアを用意しておき、各モダリティの上位N件の検索結果を保存しておく。それらの保存された検索結果同士を関連性の高い順に再ランクしてモデルに提供する手法である。この手法では再ランクをどのように行うかが重要であり、再ランクの手法が研究されている[32][33]。

本研究においては、新たな事故事例を容易に追加できるようなモジュール型の機構にすることが重要である。そこで本研究では、一般的な画像-テキストデータを学習データとしながらも、ゼロショットのデータに対しても高い性能を持つ画像-言語の共通空間モデルCLIPを使用した共通空間埋め込み型の手法でマルチモーダルRAGを構築した。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事前に生成した。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で計算する。子どもが写った画像と、子どもの事故とは無関係の画像を入力した際にCLIPが算出する全状況文との類似度スコアの分布の一例は図1に示した通りであり、事故に関係のない画像が入力された場合は類似度スコアが全体的に低くなることが分かった。そこで、類似度スコアの閾値を設定し、閾値を超える類似度の状況文が存在しない場合は画像が事故に関係ないものとして、事故事例の情報は付与しないようにした。閾値以上の類似度がある

状況文が存在する場合は子どもの事故が起り得る状況の写真と判定し、類似度が上位5件の事例の情報をLMM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、画像と警告文生成を指示するプロンプトに過去の事故の情報を追加で入力し、起りそうな事故の予測及び警告文の生成を行うRAGシステムを開発した。

LMMは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。一般的に事前学習済み大規模マルチモーダルモデルとしては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、子どもの写真や家庭内で撮影した写真をクラウド上にアップロードするのは保護者にとって心理的な抵抗感があると考えられる。そこでRAG危険警告システムの構築には、ローカル環境にダウンロードして使用できるLLaVAを用いた。

## 4. 実験

### 4.1 研究課題と仮説

本研究はLMMを活用することで、専門知識のない保護者に子どもの事故に関する知識を提供するツールを構築することを目的としている。そのため、保護者が作成するよりも適切な警告文を生成できるシステムを構築する必要がある。したがって研究課題は以下の通り。

研究課題：専門知識のない人間よりも適切性の高い警告文を生成するシステムを構築する

また、システムをRAGにすることで実際の事故に基づいた出力生成を行うため、LMMに画像のみを入力した場合よりも適切な警告文が得られると考えられる。さらに、外部データとして与えている事故データベースには多数の種類の事故が含まれているため、RAGにすることで、専門知識のない人間には予測できないような事故を予測できる(知識を補完できる)可能性が高まると考えられる。したがって、本研究の仮説は以下の2つである。

仮説1: 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

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、警告文の補完性は高くなる。

### 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。

**LMMシステム**: LMMが入力画像のみを用いて警告文を生成するシステム

**RAGシステム**: LMMが入力画像と類似事故の情報を用いて警告文を生成するシステム

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.

両システムの構築においてLLaVAの背景設定(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。背景設定は子どもの様々な日常生活の場面を写した写真を用いて繰り返し試行し、最も事故予測に適するものを検討して作成した。LLaVAは英語の会話データを学習したモデルであるため、背景設定は英語で与えた。与えた背景設定は表2に示す通りである。保育士と保護者というロールを与えた上で質問することで、LLaVAの出力がより予測事故の発生経緯について詳細に記述するようになった。更に、プロンプトの1文目に"Look at every inch of what is in this picture."という文言を入れる事で、子どもの周辺の環境や物体、状況に適した警告内容が出力されるようになった。

事事故例の検索に用いるCLIPはLLaVAに合わせ英語で事前学習済みのものを用いた。事故データベース内の各事例から事故の発生場所と発生直前にとっていた行動の情報を抜き出し、事故発生時の状況を示した文を自動生成した。その文を英語に翻訳し、CLIPを用いて入力画像との類似度検索を行った。入力画像との類似度が上位5件の事例を類似事故としてプロンプトの末尾に与えた。

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、LMMシステムとRAG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。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。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。さらに、被験者自身にも写真に対する警告文を作成させた。したがって本実験で比較する条件は以下の3条件である。

**被験者条件**: 被験者に写真に対する警告文を作成させる条件

**LMM条件**: LMMに写真のみを与えて警告文を生成させ



#### 被験者条件

感電する恐れがあるので、コンセントカバーを付けてください。また、おもちゃに破損がないかよく確認してから遊ばせてください。

#### LMM条件

画像内の子供は壁の電灯スイッチに手を伸ばしていますが、子供が遊んでいるときに誤って電灯をつけたり消したりすると、バランスを崩して転倒するなど事故につながる可能性があります。このような事故を防ぐために、保護者は電灯スイッチが子供の手の届かない高さに設置されていることを確認するか、スイッチを保護する安全カバーの使用を検討してください。

#### RAG条件

画像の子供はコンセントに手を伸ばしていますが、これは感電の重大なリスクを伴います。事故を防ぐために、親はコンセントが安全キャップで覆われているか、子供が触れないカバーが取り付けられていることを確認する必要があります。さらに、親は子供にコンセントの危険性について教育し、子供がコンセントに触れたり、物を挿入したりしないように注意深く監視する必要があります。



#### 被験者条件

机に頭を打つと危険ですので、縁や角にクッションを付けましょう。

#### LMM条件

画像に写っている子供は椅子の上に立っていますが、椅子はそのような用途には設計されておらず、転倒の大きなリスクがあります。子供が裸足で椅子の表面に立っていると、滑る可能性も高くなります。事故を防ぐために、親は子供がそのような活動用に設計されていない家具に登らないようにする必要があります。さらに、子供が登る可能性のある椅子や家具に滑り止めの表面を用意すると、滑るリスクを軽減できます。

#### RAG条件

画像の子供はテーブルの上に立っており、転倒して怪我をする大きなリスクがあります。親は常に子供を監視し、家具の上に立つなどの潜在的に危険な活動に従事しないようにすることが重要です。

図2 3条件で得られた警告文の例

る条件

**RAG条件**：LMMに写真とそれに関する事象例を与えて警告文を生成させる条件

それぞれの条件で得られた警告文を図2に示す。ただし、被験者条件で得られた警告文の例としては、LMM条件、RAG条件よりも自身が書いた警告文の方が優れていると評価する回数が多かった被験者の警告文を掲載している。

次に、3つの条件で作成された警告文に対して被験者アンケートを行い、警告文の適切性を評価させた。被験者条件とシステムを用いた2条件、LMM条件とRAG条件の間で警告文の適切性の評価に差があるかを調査した。また、LMM条件とRAG条件で生成された警告文の内容の補完性についての評価項目を設け、両システムの補完性に差があるか調べた。

#### 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について、作成直後は客観的に評価できない可能性があるため、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で被験者本人が作った警告文、LMMシステムが生成した警告文、RAG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、LMMとRAG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。警告文について、適切に危険を警告できているか(適切性)、被験者が気づけなかった危険をシステムが指摘できているか(補完性)について評価させた。この2つ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。

危険警告システムの利用者は子育て経験が乏しい保護者であるため、子育て経験のない若者を被験者として選出した。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。評価順による回答への影響を除去するため、被験者ごとに異なる順番で写真を提示した。

また、適切性の評価については信憑性を裏付けるため、子育ての専門知識を有した者にもLMM条件とRAG条件の警告文の適切性を評価させた。専門家として、保育児童学の研究者にアンケートを行った。

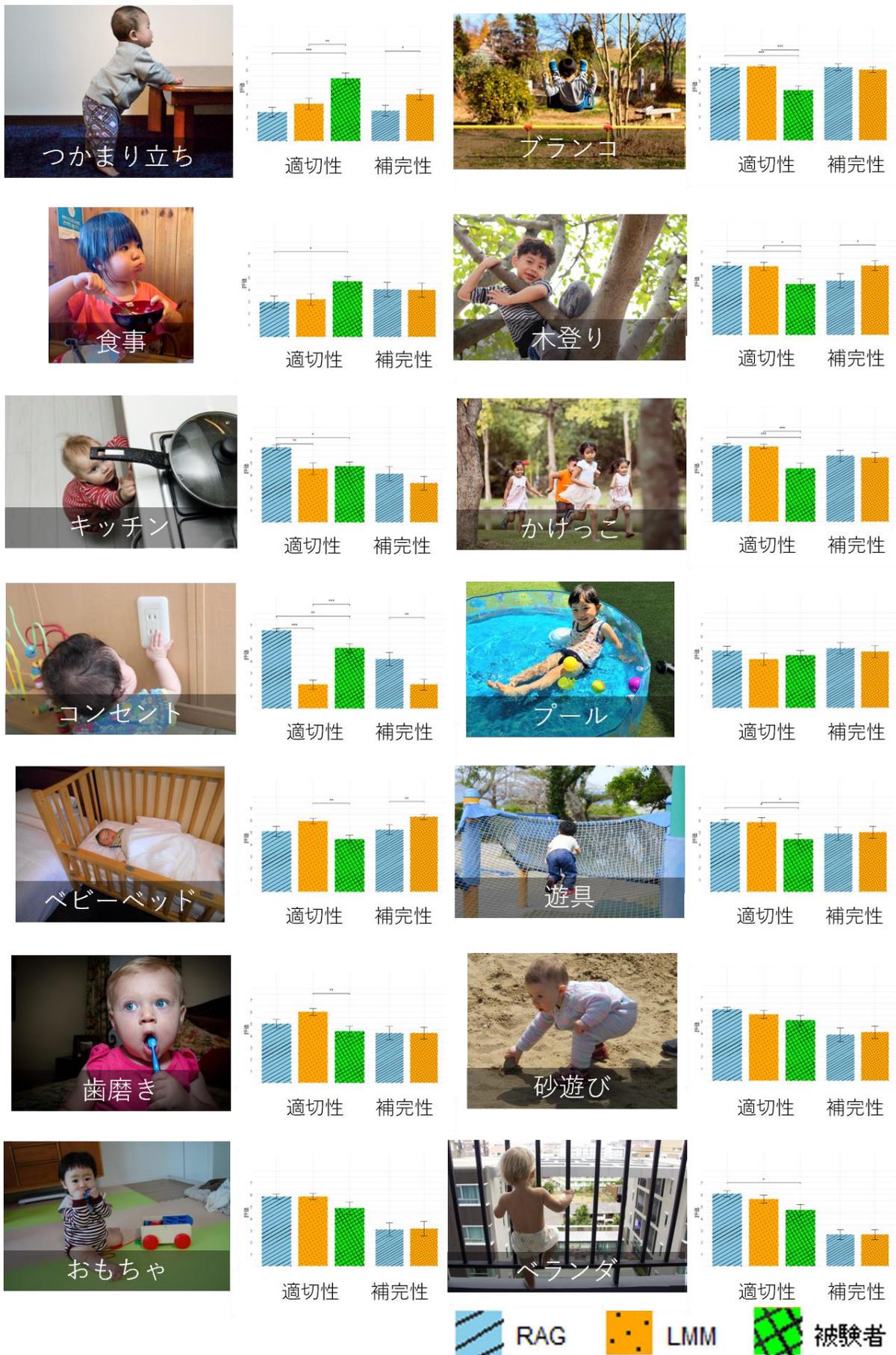


図3 各写真の評価結果

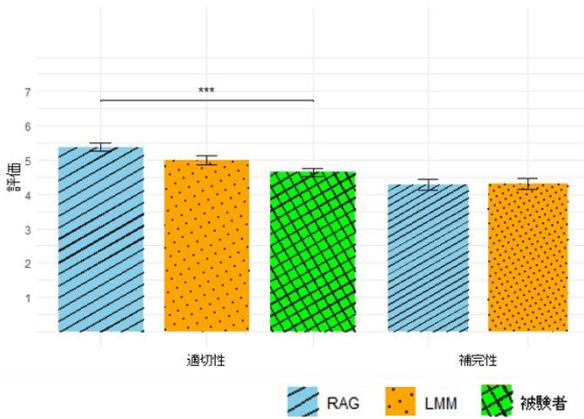


図 4 写真全体の評価結果

## 5. 実験結果

### 5.1 適切性の評価結果

14 枚の写真に対して、被験者条件、LMM 条件、RAG 条件それぞれで作った警告文の適切性の評価平均の比較を図 3 に、画像全体の評価平均の比較を図 4 に示す。図 3 では、写真ごとに適切性（左）と補完性（右）の 2 種類の評価結果を並べて表示している。左側の 3 本の棒グラフは適切性についての 3 条件（RAG 条件、LMM 条件、被験者条件）の評価平均値と標準誤差を示したものである。凡例に示す色は条件を表し、グラフ上部の有意差バーは多重比較により有意差が確認された条件間を示している。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、つかまり立ちの画像では被験者条件が LMM 条件・RAG 条件よりも適切性が有意に高く ( $p<0.01$ )、食事の画像では被験者条件が RAG 条件よりも有意に高かった ( $p<0.05$ )。また、ブランコ・木登り・かけっこ・遊具の画像では LMM 条件・RAG 条件が被験者条件に比べて有意に高かった ( $p<0.05$ )。キッチンの画像では、RAG 条件が被験者条件・LMM 条件に比べて有意に高かった ( $p<0.05$ )。コンセントの画像では RAG 条件が被験者条件・LMM 条件よりも有意に高く ( $p<0.01$ )、被験者条件が LMM 条件よりも有意に高かった ( $p<0.001$ )。ベビーベッド・歯磨きの画像では LMM 条件が被験者条件よりも有意に高く ( $p<0.01$ )、ベランダの画像では RAG 条件が被験者条件よりも有意に高かった ( $p<0.05$ )。

### 5.2 補完性の評価結果

14 枚の写真に対して、LMM 条件と RAG 条件で生成された警告文の補完性の画像毎の評価平均の比較を図 3 に、画像全体の評価平均の比較を図 4 に示す。図 3 では、写真ごと適切性（左）と補完性（右）の 2 種類の評価結果を並べて表示しており、右側の 2 本の棒グラフは補完性についての 2 条件（RAG 条件、LMM 条件）の評価平均値と標準誤差を示したものである。有意差につ

表 3 子育て未経験者と専門家の評価のクロス集計表

(a) LMM 条件

		子育て未経験者	
		高評価	低評価
専門家	高評価	4	0
	低評価	7	3

(b) RAG 条件

		子育て未経験者	
		高評価	低評価
専門家	高評価	9	0
	低評価	3	2

いて t 検定を行った結果、つかまり立ち・木登り・ベビーベッドの画像では LMM 条件が RAG 条件よりも値が有意に高かった ( $p<0.05$ )。また、コンセントの画像では RAG 条件が LMM 条件よりも有意に高かった ( $p<0.01$ )。

### 5.3 専門家による評価結果

専門家に、14 枚の画像に対して LMM 条件および RAG 条件で得られた警告文の適切性を評価させた。それぞれの条件における子育て未経験者の評価平均と専門家の評価の比較結果を表 3 に示す。評価は 7 段階のリッカート尺度で行われたものであり、「5. ややあてはまる」以上を高評価、「4. どちらともいえない」以下を低評価として 2 値化し、クロス集計を行った。いずれの条件においても、子育て未経験者が低評価とした写真に対して、専門家が高評価を与えるという評価の相違は確認されなかった。また、両条件ともに、子育て未経験者と専門家の両方から低評価を得る例は少なく、ほとんどの例で子育て未経験者か専門家の少なくともどちらか一方からは高評価を得ていた。さらに、LMM 条件と比較して、RAG 条件では子育て未経験者と専門家のいずれからも高評価を得た写真の数が倍以上多かった。一方、子育て未経験者からは高評価を得た写真に専門家が低評価を与えた例は LMM 条件の方が RAG 条件よりも多かった。LMM 条件の方が素人と専門家で評価がわかれやすかった理由については 6 節で詳しく考察する。

## 6. 考察

### 6.1 被験者とシステムの警告文の適切性の比較

適切性の評価では、14 枚中 6 枚の写真について LMM 条件の方が被験者条件よりも適切であり、14 枚中 8 枚の写真について RAG 条件の方が被験者条件よりも適切であることが分かった。LMM 条件と RAG 条件の両方の警告文が被験者条件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 条件の方が被験者条件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誤認識が起っており、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誤認識したりしていた。このように、学習データとして使われた画像と実験に用いた写真の文化の差や、誤認識が原因で、一部の写真ではシステムは被験者条件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、被験者条件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、LMM を使用することで、人間が作成する警告文よりも適切な警告文を生成できるシステムを構築するという研究課題を達成できた場合もあった。

## 6.2 事件事例の有無による警告文の適切性の比較

14 枚中 2 枚の写真で LMM 条件よりも RAG 条件の警告文が適切だと評価されており、LMM 条件の警告文が RAG 条件よりも適切だと評価された写真は存在しなかった。RAG 条件が優れた評価を得た 2 枚の写真では、LMM 条件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 条件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んでいたので、誤認識が起らず、状況に見合った警告文を生成できていた。このように、RAG 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 1 の通りの結果が得られた画像もあった。

## 6.3 補完性の評価結果

補完性の評価では、LMM 条件の方が高く評価された写真が 14 枚中 3 枚あり、LMM 条件の方が被験者の知識を補完するのに優れている可能性が高い。このような評価にした理由を被験者に質問したところ、RAG 条件の警告文はよく起こる事故について警告しており補完性がなかったが、LMM 条件は被験者が思いつかない事故について警告していたため、LMM 条件の補完性が高くなっていることがわかった。RAG 条件は実際に起こった事故をもとに警告文を生成しているため、起こりやすい事故に関する警告文になっており、最も気を付けるべきことについて警告できる一方で被験者が思いつけない警告をできることは少なく、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補完性は高くなると予測した仮説 2 とは異なる結果であった。

## 6.4 子育て未経験者と専門家の評価の比較

事故予測と対策方法が共に適切であるとして子育て未経験と専門家の両方から支持された警告文は RAG 条件の方が LMM 条件よりも多かった。その一方で、LMM 条件では子育て未経験者からは高評価を得た写真に専門家が低評価を与える例が多く、RAG 条件では少数の例にとどまっていた。専門家と子育て未経験者の評価基準の違いとして、子育て未経験者は発生確率が低い事故であっても起こりえる事故が予測できていれば高評価をつける一方で、専門家の観点では起こることが稀な事故よりも発生確率の可能性が高い事故を優先して警告することが適切だと評価するポイントであることがわかった。RAG 条件は発生確率が高い事故を優先的に警告していた一方、LMM 条件は気づきにくい事故に関する警告が多かったため、子育て未経験者から高評価を得られても専門家からは低評価をつけられる例が多かったのだと考えられる。さらに、保護者がとるべき行動が明確に提示されている警告文が専門家から高く評価される一方で、家具や遊具等の設備の安全点検といった、警告されても保護者が自力で解決しにくい内容を事故防止対策として提示している警告文は専門家から低評価をつけられやすい傾向にあることもわかった。また、専門家は事故による怪我の重大性も評価の上で考慮していた点の子育て未経験者と異なっていた。この 2 点においても、LMM 条件では、家具や設備の安全点検を促す警告や軽微な事故に関する警告が多く得られたのに対し、RAG 条件では、子どもの行動面に着目した警告や、重篤な事故に繋がるおそれのある状況を優先的に警告する傾向が見られた。これにより、RAG 条件にすることでより専門家の評価基準に整合した警告文が得やすくなったと考えられる。

## 6.5 今後の展望

本システムは複数のモジュールとデータベースを組み合わせるため、入れ替えや追加が容易である。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、事故データベース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、最新の事件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あるため、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。

また、評価項目を増やすことでより多く観点からシステムを評価する必要があると考える。子育て現場において、事故の種類は非常に多い。子どもの死亡事故や障害が残る事故を防止するためには、重大な事故ほど優先して予測・防止する必要がある。今回は警告文の適切性と補完

性のみに着目して実験を行ったが、今後は重大な事故を優先的に警告できていたか(重大性)に関しても高い評価が得られるシステムの構築手法についても検討する必要がある。

本研究では親の知識獲得を目的としリアルタイム性を考慮しなかったが、将来的にはリアルタイムに警告文を生成することでより子育て現場に貢献できると考えられる。子どもが過ごす場所に見守りカメラを設置し、カメラから得られた画像に対してリアルタイムに事故予測・警告文生成を行う事ができれば、周囲の大人が子どもから目を離れた間に発生する事故に対する防止策としてシステムを使用することができ、保護者の負担を大幅に軽減することができる。本研究でもシステムの高速化についても研究したが、依然リアルタイムシステムとして運用するのは難しい。本システムは画像を入力してから警告文を得られるまでにおよそ 40~50 秒程度かかり、警告文が長い場合は 50 秒を超える。システムの稼働にかかる時間のほとんどは大規模マルチモーダルモデルが出力を生成するのにかかる時間である。子どもの事故は短い時間の間に発生し、転倒事故については転び始めてから 0.5 秒で転ぶという調査結果もある。このような事故にも対応できるよう、遅くとも 0.5 秒以内に警告できるように警告文生成部を高速化する必要がある。大規模言語モデルは 1 トークンずつ逐次的に生成する特徴をもつため、出力が長いほど生成にかかる時間も長くなる。大規模マルチモーダルモデルの高速化に加えて、緊急性に応じて警告文の長さを調節するようなシステムを開発することが効果的である可能性がある。

## 7. まとめ

子どもの事故を回避する様々なシステムが提案されてきた。そのようなシステムは、事故が起こる過程をシミュレートしたり、事故に関する知識を保護者に提示したり、特定の事故や一般的な事故を対象としているものがほとんどである。これに対し、マルチモーダル大規模言語モデル(LMM)を活用することで、子どもの育児環境で起こり得るより多様な事故を予測・警告することができると考えた。本研究は、LMM で子どもの事故を警告する次の 2 種類のプロトタイプシステムを構築し、育児の未経験者と専門家を対象とした評価実験を通してその設計指針を示すものである。システム A: 入力した育児環境の写真とプロンプトに基づいて起こり得る事故と回避方法を説明する警告文を出力する。システム B: 入力写真の状況から子どもの事故データベースを検索し、プロンプトと共に事故事例も入力する。これらのシステムに 14 枚の写真を入力して得られた警告文の適切性と補完性を育児の未経験者が評価した結果、システム B の警告文の適切性が高く評価される傾向が見られたが、システム A において補

完性が高く評価される場合があった。専門家の評価と聞き取り調査によると、起こることが稀な事故よりも可能性が高い事故を優先して警告することが適切であり、それ故に過去の事故事例に基づいているシステム B の警告文が専門家からも高く評価されやすいことが分かった。したがって、子どもの事故を警告する LMM システムの設計においては、まず育児環境に合致した過去の事後事例を検索・入力して得られた警告を提示し、さらに事故事例を入力せずに得られたその環境特有の稀な事故の警告を提示することが有益であると考えられる。環境特有の事故を発見するように LMM に入力写真を注目させるためには、プロンプトにおいて”Look at every inch of what is in this picture.”というフレーズ等を用いることが有効であった。これらの知見により、子どもの危険回避に大規模言語モデルを活用する研究が促進されることを期待する。

**謝辞** 本研究は、JSPS 科研費 JP22K12126, JP24K00066, JP25K15255, JST 次世代研究者挑戦的研究プログラム JPMJSP2107 の支援を受けたものです。子どもの事故データベースを提供して下さった国立成育医療研究センターに深謝の意を表します。助言をいただいた駒沢女子短期大学の猪熊弘子教授、京都工芸繊維大学深田智教授に感謝いたします。

## 参考文献

- [1] “子どもの事故”. [https://www.kokusen.go.jp/soudan\\_now/data/kodomo\\_jiko.html](https://www.kokusen.go.jp/soudan_now/data/kodomo_jiko.html) (参照 2025-5-1).
- [2] Rhodes, A., Kheireddine, S., and Smith, A. D.. Experiences, Attitudes, and Needs of Users of a Pregnancy and Parenting App (Baby Buddy) During the COVID-19 Pandemic: Mixed Methods Study. *JMIR mHealth and uHealth*, 2020, vol.8, no.12, e23157.
- [3] “乳幼児期の事故防止学習ソフト”. <https://www.fukushi.metro.tokyo.lg.jp/> (参照 2025-10-1).
- [4] 二又川 求哉, 牧野 哲裕, 田中 直登, 正 小林 宏, 橋本 卓弥, 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*, 2018, 2P2-B16.
- [5] Simo, A., Nishida, Y., and Nagashima, K.. A Humanoid Robot to Prevent Children Accidents. *Interactive Technologies and Sociotechnical Systems*, 2006, pp. 476-485.
- [6] Abe, K., Shiomi, M., Pei, Y., Zhang, T., Ikeda, N., and Nagai, T.. ChiCaRo: tele-presence robot for interacting with babies and toddlers. *Advanced Robotics*, 2018, vol.32, no.4, pp. 176-190.
- [7] “遠隔育児支援ロボット ChiCaRo”. <http://www.rlg.sys.es.osaka-u.ac.jp/chicaro/> (参照 2024-10-28).
- [8] “VisionPose 活用事例：見守り分野”. <https://www.next-system.com/visionpose/use/watching> (参照 2025-2-5).
- [9] Sundarajoo, R. A., Chung, G. C., Pang, W. L., and Tan, S. F.. A Remote Baby Surveillance System with RFID and GPS Tracking. *arXiv preprint arXiv:2211.14553*, 2022.
- [10] Khan, T.. An Intelligent Baby Monitor with Automatic Sleeping Posture Detection and Notification. *AI*, 2021, vol. 2, no. 2, pp. 290-306.
- [11] Nose, T., Kitamura, K., Oono, M., Nishida, Y., and Ohkura, M.. 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*, 2020, vol.11, pp. 5845-5855.
- [12] Kitamura, K., Nishida, Y., Matsumoto, N., Motomura, Y., Yamanaka, T., and Mizoguchi, H.. Development of Infant Behavior Simulator: Modeling Grasping Achievement Based on Developmental Behavior Model and Environmental Interest Induction Model. *Journal of Robotics and Mechatronics*, 2005, vol.17, no.6, pp. 705-715.
- [13] Nishizaki, M.. Visualizing affordances of everyday objects using mobile augmented reality to promote safer and more flexible home environments for infants. *Advances in Science, Technology and Engineering Systems Journal*, 2018, vol.3, no.6, pp. 74-82.
- [14] Achiam, J., et al.. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.
- [15] Georgiev, P., et al.. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530, 2024.
- [16] Dai, W., Lee, N., Wang, B., Yang, Z., Liu, Z., Barker, J., Rintamaki, T., Shoeybi, M., Catanzaro, B., and Ping, W.. NVLM: Open Frontier-Class Multimodal LLMs. arXiv preprint arXiv:2409.11402, 2024.
- [17] Liu, H., Li, C., Wu, Q., and Lee, Y. J.. Visual instruction tuning. arXiv preprint arXiv:2304.08485, 2023.
- [18] Singhal, K., et al.. Toward expert-level medical question answering with large language models. *Nature Medicine*, 2025, vol.1, no.8.
- [19] Chen, L., et al.. PharmaGPT: Domain-specific large language models for bio-pharmaceutical and chemistry. arXiv preprint arXiv:2406.18045, 2024.
- [20] Yang, H., Liu, X. Y., and Wang, C. D.. FinGPT: Open-Source Financial Large Language Models. arXiv preprint arXiv:2306.06031, 2023.
- [21] “Palmyra-Fin-70B-32k: a powerful LLM designed for Finance”. <https://dev.writer.com> (参照 2025-5-27).
- [22] Zhou, Z., Shi, J. X., Song, P. X., Yang, X. W., Jin, Y. X., Guo, L. Z., and Li, Y. F.. Lawgpt: A chinese legal knowledge-enhanced large language model. arXiv preprint arXiv:2406.04614, 2024.
- [23] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.-T., and Rocktaschel, T.. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 9459–9474.
- [24] Ovadia, O., Brief, M., Mishaeli, M., and Elisha, O.. Fine-tuning or retrieval? comparing knowledge injection in llms. arXiv preprint arXiv:2312.05934, 2023.
- [25] Soudani, H., Kanoulas, E., and Hasibi, F.. Fine tuning vs. retrieval augmented generation for less popular knowledge. *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, 2024, pp. 12-22.
- [26] Chen, W., Hu, H., Chen, X., Verga, P., and Cohen, W. W.. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. arXiv preprint arXiv:2210.02928, 2024.
- [27] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I.. Learning Transferable Visual Models From Natural Language Super. *Proceedings of the 38th International Conference on Machine Learning*, 2021, pp. 8748-8763, PMLR 139:8748-8763.
- [28] Lyu, Y., Zheng, X., Zhou, J., and Wang, L.. Unibind: Llm-augmented unified and balanced representation space to bind them all. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26752-26762.
- [29] Awadalla, A., et al.. Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390, 2023.
- [30] Lin, W., and Byrne, B.. Retrieval augmented visual question answering with outside knowledge. arXiv preprint arXiv:2210.03809, 2022.
- [31] Li, J., Li, D., Savarese, S., and Hoi, S.. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International conference on machine learning*, 2023. pp. 19730-19742, PMLR.
- [32] Wen, H., Zhuang, H., Zamani, H., Hauptmann, A., and Bendersky, M.. Multimodal reranking for knowledge-intensive visual question answering. arXiv preprint arXiv:2407.12277, 2024.
- [33] Chen, Z., Xu, C., Qi, Y. and Guo, J.. LMM Is a Strong Reranker: Advancing Multimodal Retrieval-augmented Generation via Knowledge-enhanced Reranking and Noise-injected Training. arXiv preprint arXiv:2407.21439, 2024.

## 付録

### 付録 A. 使用画像の出典一覧

本研究の実験で利用した画像 14 枚の出典を以下に示す。いずれも著作権フリーの画像提供サイトから取得したものであり、研究利用に問題がないことを確認している。

- “PhotoAC” <https://www.photo-ac.com> : 図 2(b), 図 3 (つかまり立ち), 図 3 (食事), 図 2(a), 図 3 (コンセント), 図 3 (プール), 図 3 (ベビーベッド), 図 3 (遊具), 図 3 (おもちゃ)
- “Pixabay” <https://pixabay.com/ja/> : 図 3 (ブランコ), 図 3 (木登り), 図 3 (歯磨き), 図 3 (砂遊び)
- “iStock” <https://www.istockphoto.com/jp> : 図 3 (キッチン)
- “Dreamstime” <https://www.dreamstime.com/> : 図 3 (かけっこ), 図 3 (ベランダ)