時間とともに変化する未整理な外部知識の継続的な整理と長文コンテキスト RAG への活用

西田典起♠　Fei Cheng♢　松本裕治♠
♠理化学研究所　♢京都大学
noriki.nishida@riken.jp

概要

本研究では，外部知識をタイムスタンプ付き命題の集合として表現し，命題間の時間的更新関係や矛盾関係，根拠関係を明示的に構造化する RAG フレームワーク，Temporally-structured Proposition Relations RAG（TPR-RAG）を提案する．TPR-RAG は，質問に応じて関連命題とその関係構造を検索し，LLM が時間的妥当性や情報の信頼性を考慮して推論できる構造化コンテキストを生成する．実験により，TPR-RAG は従来のチャンクベースおよび命題ベースの RAG 手法と比べて，頑健性と精度の両面で一貫した性能向上を示すことを確認した．

1　はじめに

大規模言語モデル（LLM）の内部知識は時間とともに陳腐化する．この問題に対して，継続的なファインチューニングや局所的なパラメータ編集が提案されてきたが [1, 2]，これらの手法は更新を繰り返すと破滅的忘却が生じやすく，知識の一貫性を長期的に維持することは難しい [3, 4]．

この問題への有力な対処法として，検索拡張生成（Retrieval-Augmented Generation; RAG）が広く用いられている [5]．しかし，外部知識もまた時間とともに変化し，しばしば未整理であるため，時間的に古くなった記述や相互に矛盾する主張，信頼性の低い情報が混在する．このようなコンテキストに対して，LLM が頑健に推論できないことが報告されている [6, 7]．

検索精度を高めるため，文書をより細かな事実単位である 命題 に分解する RAG が提案されている [8]．しかし，多くの命題ベース RAG は命題を独立に扱い，命題間の更新関係や矛盾関係を明示的にモデル化していないため，時間更新・矛盾・誤情報を含む知識に対して依然として脆弱である [6, 7]．

この問題は，長文コンテキスト RAG [9] において特に顕著になる [10]．コンテキスト長が拡大すると，時間更新・矛盾・誤情報を含む多数の命題が同時に参照され，整合しない情報が同一コンテキスト内に混在する．命題間の関係を明示的に扱わない場合，長文 RAG はこれらをすべて同時に妥当な情報として扱ってしまい，推論が不安定になる．そこで，非構造化された命題列からモデル自身に更新のトレースや矛盾の解消をさせるのではなく，それらを保持・明示する構造的なガードレールが必要となる．

本論文では，時間とともに変化する未整理な外部知識を，長文 RAG のために継続的に整理・活用するフレームワーク **TPR-RAG** (Temporally-structured Proposition Relations for RAG) を提案する．TPR-RAG は，外部知識をタイムスタンプ付き命題の集合として表現し，「更新」「矛盾」「根拠」を表す命題間関係をグラフ構造として整理する．実験により，TPR-RAG は既存のチャンクベース，命題ベース，時間情報付き命題ベースの RAG と比べて，高い精度と頑健性を示すことを確認した．この結果は，時間更新・矛盾・誤情報を含む外部知識に対して，命題間関係を明示的に扱うことの有効性を示唆する．

2　手法

本研究では，時間とともに変化する未整理な外部知識を長文 RAG で扱うためのフレームワーク **TPR-RAG** を提案する．システム全体の構成を図 1 に示す．TPR-RAG は，文書から抽出したタイムスタンプ付き命題をノードとし，命題間の更新・矛盾・根拠関係をエッジとして表す命題グラフを構築する．本フレームワークは，命題抽出，命題間関係抽出，部分グラフ検索，回答生成の 4 段階からなり，前半 2 段階で文書集合から命題グラフを構築し，後

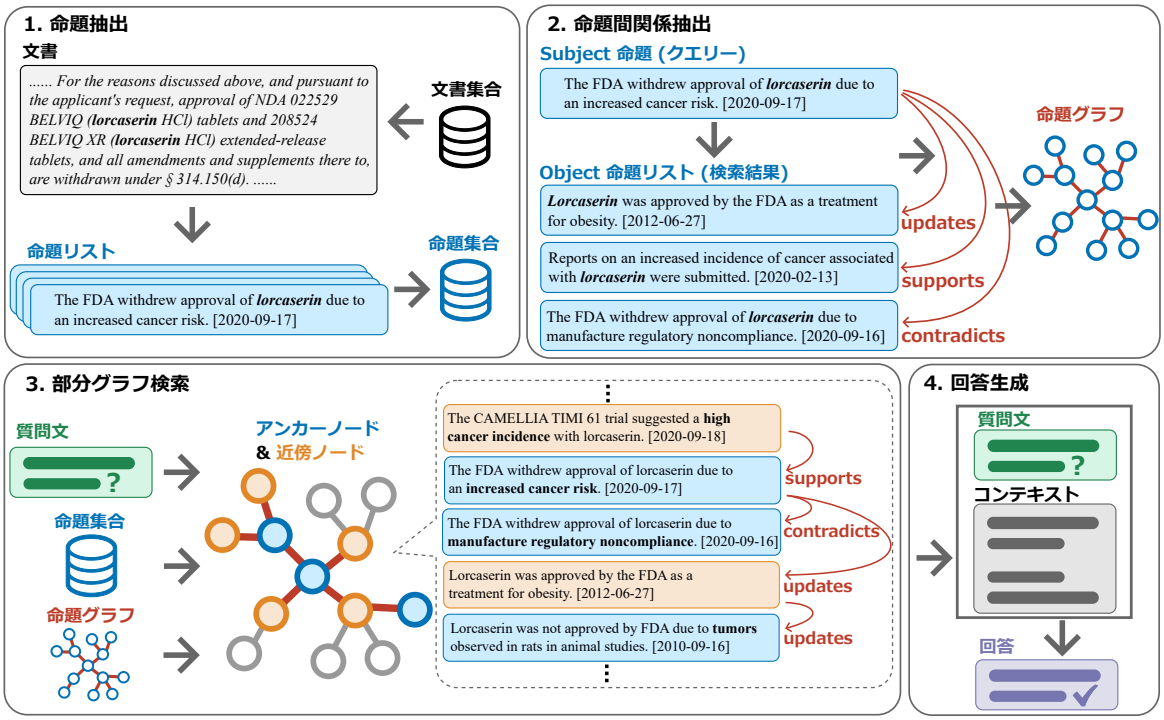

図 1: 提案手法 TPR-RAG の概要．TPR-RAG は，文書をタイムスタンプ付き命題の集合として表現し，命題間の更新・矛盾・根拠関係を明示的にモデル化することで，時間とともに変化する未整理な外部知識を長文 RAG のために継続的に整理する．

半 2 段階で質問に応じて検索と回答生成を行う．

命題抽出　本研究では，知識の最小単位として，ある時点における事実を表す **命題** を用いる．各命題は簡潔かつ自己完結的であり，文脈依存の表現は事前に解消され，`YYYY-MM-DD` 形式のタイムスタンプが付与される．生テキストからの命題抽出には LLM（GPT-4o-mini [11]）を用いて，各文書から複数の命題を抽出する．得られた命題は単一の命題集合として集約され，後続の命題間関係抽出および部分グラフ検索で用いられる．

命題間関係抽出　命題を独立した単位として扱うことの限界を克服するため，本研究では命題間の関係を明示的にモデル化する．各関係は時間制約付きの有向関係として定義され，時間的に新しい命題（subject）が，過去の命題（object）を更新，否定，または支持するという仮定に基づく．具体的には，`updates`，`contradicts`，`supports`, `NOREL` (関係なし) の 4 タイプを用いる．命題集合が与えられたとき，各命題をクエリとして，意味的に関連し，かつ時間的に過去の命題 20 件を Contriever [12] を用いて検索し，相対的に軽量な LLM（GPT-4o-mini）を用いて命題間関係を分類する．命題間関係の品質は下流の推論性能にとって極めて重要であるため，推論時に実際に使用されるエッジのみを対象として，より高性能な LLM（GPT-4o）により再評価を行う精緻化ステップを導入する．最終的に，命題をノード，命題間関係をエッジとする有向命題グラフ $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ を構築する．

部分グラフ検索　推論時には，ユーザーの質問に直接関連する命題だけでなく，それらがどのように更新・矛盾・支持しあうかという関係構造も取得する．まず，質問 q に対して Contriever による意味的類似度に基づき，上位 k 件の関連命題を検索し，質問内容と直接対応する事実を取得する．次に，これらの命題をグラフ上のアンカーノードとして，入出力エッジで直接接続される近傍命題を取得する．最後に，取得した命題集合（アンカー命題+近傍命題）に含まれる命題間関係をすべて収集し，質問に関連する命題とその関係構造からなる部分グラフ $\mathcal{G}^q = (\mathcal{V}^q, \mathcal{E}^q)$ を構成する．

回答生成　検索された部分グラフ $\mathcal{G}^q = (\mathcal{V}^q, \mathcal{E}^q)$ を，QA モデルが利用可能なテキストコンテキストへ変換する．具体的には，命題をタイムスタンプ順（昇順）に整列し，各命題を時刻と事実記述からなる項目として言語化するとともに，命題間の更新・矛盾・支持関係を説明文として付与する

手法	Asteria			CLARK-News			平均
	k=10	k=100	k=400 (all)	k=10	k=100	k=1000	
GPT-4o (no context)	36.7	36.7	36.7	34.7	34.7	34.7	35.7
+ チャンクベース RAG	-	-	-	54.9	53.9	49.2	52.7
+ 命題ベース RAG	37.8	36.1	41.1	44.6	54.4	47.7	43.6
+ 時間情報付き命題ベース RAG	74.4	65.6	58.9	58.5	84.5	76.2	69.7
+ TPR-RAG (ours)	82.8	88.3	**86.1**	54.9	**85.0**	**82.4**	79.9
+ 近傍拡張	**87.2**	**89.4**	**86.1**	**70.5**	**85.0**	*	**83.6**
Llama-3.1-70B-Instruct (no context)	35.0	35.0	35.0	34.2	34.2	34.2	34.6
+ チャンクベース RAG	-	-	-	47.7	36.8	46.6	43.7
+ 命題ベース RAG	36.1	31.7	31.7	37.3	45.6	35.8	36.3
+ 時間情報付き命題ベース RAG	71.1	64.4	54.4	51.3	77.7	49.2	61.4
+ TPR-RAG (ours)	79.4	**76.1**	**68.3**	53.9	81.9	**75.1**	72.5
+ 近傍拡張	**87.8**	75.6	**68.3**	**66.8**	**82.4**	*	**76.2**

表 1: 時間的多肢選択 QA ベンチマークにおける正解率（%）．k は検索される事実（チャンクまたは命題）の数を表し，アスタリスク（*）はコンテキスト長制限により評価できなかった結果を示す．

ことで，時間的推移と関係構造が読み取りやすい形で提示する．言語化された命題の例を付録の図 3 に示す．得られたテキストコンテキストを $\mathcal{C}$ = Contextualize($\mathcal{V}^q, \mathcal{E}^q$) と表す．質問 q とコンテキスト $\mathcal{C}$ が与えられたとき，QA モデル（GPT-4o または Llama-3.1-70B-Instruct [13]）はこれらを入力として最終的な回答 a = QA($q, \mathcal{C}$) を生成する．

3 実験

実験設定 時間的質問応答ベンチマークとして，構築したデータセット **Asteria** と，実世界ベンチマークである **CLARK-News** [14] を用いて評価を行う．両データセットでは，すべての質問が明示的な時間情報とともに “`{question} (Date: {timestamp})`” の形式で与えられる．**Asteria** は，時間更新・矛盾・誤情報を含む未整理な外部知識に対する推論能力を評価するために構築した時間的 QA データセットである．各質問は時間条件に応じて正解が変化する多肢選択形式で表現され，180 件のユニークな時間的質問応答ペアを含む．詳細な構築手順および例は付録 B に示す．**CLARK-News** は，ニュース記事に基づく実世界の時間的 QA ベンチマークである [14]．同一質問文で正解が時間とともに少なくとも 3 回変化する事例のみを残し，すべての質問を多肢選択形式へ変換する．前処理後，193 件のユニークな時間的質問応答ペアを含む．

すべてのデータセットは多肢選択 QA として定式化されているため，評価指標には正解率を用いる．選択された回答が，指定された時間条件下での正解選択肢と一致する場合に正解と判定する．

実験では以下の手法との比較を行う．(1) **LLM（no context）**：外部検索を行わず，言語モデル単体で回答する．(2) **チャンクベース RAG** [5]：文書を固定長チャンク (100 トークン) に分割して検索する非時間的 RAG．(3) **命題ベース RAG** [8]：文書を命題に分解して検索するが，時間情報は付与しない．(4) **時間情報付き命題ベース RAG**：命題にタイムスタンプを付与し時系列順に整列するが，命題間関係は考慮しない [15]．

結果と考察 表 1 は，検索する命題数 k を変化させた場合の，Asteria および CLARK-News [14] における正解率を示す．(1) **命題間関係の明示的モデル化は有効である．** TPR-RAG は，すべてのデータセットおよび検索設定において比較手法を一貫して上回り，特に大規模な検索設定で時間情報付き命題ベース RAG を大きく上回った．この結果は，命題を独立に扱うのではなく，命題間の更新・矛盾・根拠関係を明示的にモデル化することが有効であることを示している．(2) **関係に基づく近傍拡張はさらなる性能向上をもたらす．** 初期に検索された命題に対して，関係で接続された近傍命題を追加することで，多くの設定において性能がさらに向上した．これは，初期に検索された命題集合に加えて，関係

手法	Asteria (k=400)
TPR-RAG (Full)	**86.1**
w/o 精緻化ステップ (Section 2)	83.8
w/o 関係ラベル	82.2
w/o `updates` 関係	81.7
w/o `contradicts` 関係	83.3
w/o `supports` 関係	80.6

表 2: Asteria におけるアブレーション実験の結果（k=400）．QA モデルには GPT-4o を用いる．

で接続された命題を取り込むことで，相補的な情報が提供されることを示唆している．(3) **構造がある場合に長文コンテキストは有効である．**検索サイズ k を増加させた際の影響は，検索表現に強く依存した．チャンクベース RAG では k の増加に伴い性能が低下する一方で，TPR-RAG は大規模な検索サイズにおいても高い性能を維持した．これは，時間的更新や矛盾関係を明示的に扱うことで，長文コンテキストに起因するノイズの影響を抑制できることを示唆している．(4) **時間情報は必要だが，関係構造が不可欠である．**命題分解のみでは，チャンクベース RAG に対して一貫した性能向上は得られなかった．これは，命題分解のみでは時間的手がかりを十分に保持できない可能性を示唆している．また，命題への時間情報の付与は性能を改善するものの，TPR-RAG との間には依然として大きな差が存在した．この結果は，時間指定は有効ではあるが，更新や矛盾といった命題間関係を伴わなければ十分ではないことを示している．

アブレーション実験　命題間関係の寄与を分析するため，Asteria 上で検索サイズを k=400 に固定したアブレーション実験を行う．結果を表 2 に示す．精緻化ステップを除去すると性能が明確に低下し，高品質な命題間関係が重要であることが分かった．また，グラフ構造を維持したまま関係ラベルを単一の汎用表現に置き換えると正解率が大きく低下し，関係の意味情報が有効であることが示された．さらに，いずれか一種類の関係ラベルを除去した場合にも一貫して性能が低下した．以上より，TPR-RAG の性能向上には，正確な関係抽出と命題間関係の明示的なモデル化が有効であることが確認できる．

偽命題の構造的分析　TPR-RAG により構築された命題グラフにおいて，偽命題がどのように扱われているかを分析する．Asteria では，各命題が偽または非偽としてアノテーションされており，これを用

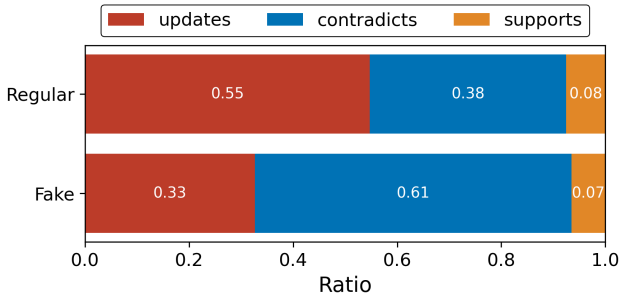

図 2: Asteria における偽命題と通常（非偽）命題に付与された関係ラベルの分布．

いて偽命題の構造的な位置づけを調べる．まず，偽命題の 57.6%が `contradicts` または `updates` 関係を通じて他の命題と接続されており，多くの偽命題が孤立せず，矛盾関係や時間的更新連鎖の中に組み込まれていることが分かった．さらに，関係ラベルの分布（図 2）を比較すると，偽命題は `contradicts` 関係で参照される割合が高く，時間的に新しい命題によって否定されやすい傾向が見られた．一方，通常命題は主に `updates` 関係で接続され，事実の時間的進展を反映した更新連鎖の一部として扱われている．

命題の情報十分性　抽出された命題集合が，質問応答に十分な情報を保持しているかを検証するため，検索された命題部分グラフに対して，関連する元文書を補助的に追加する実験を行う．各文書は，それが含むアンカー命題の数に基づいてスコア付けされ，上位 r 件の文書を補助参照文書として選択する．その結果，補助文書の追加効果は一貫しておらず，2 件追加した場合には改善（82.4% → 83.9%）が見られる一方で，1 件のみ追加した場合には性能が低下した（81.9%）．これは，コンテキスト長の増加に対して新規情報量が不十分となり，情報密度が低下したためであると考えられる．以上より，命題部分グラフは質問応答に必要な情報の大部分をすでに保持しており，元文書の追加は体系的な利点をもたらさない可能性が示唆される．

4　おわりに

本論文では，時間とともに変化する未整理な外部知識を，命題間の関係性を明示的にモデル化することで継続的に整理し，長文 RAG の推論に活用するフレームワークを提案した．実験により，時間更新・矛盾・誤情報を含む外部知識に対して，命題間関係を明示的に扱うことの有効性を確認した．

参考文献

[1] Yingming Zheng, Hanqi Li, Kai Yu, and Lu Chen. When long helps short: How context length in supervised fine-tuning affects behavior of large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 10293–10308, Suzhou, China, November 2025. Association for Computational Linguistics.

[2] Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6491–6506, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[3] Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. Time waits for no one! analysis and challenges of temporal misalignment. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 5944–5958, Seattle, United States, July 2022. Association for Computational Linguistics.

[4] Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning LLMs on new knowledge encourage hallucinations? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 7765–7784, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[5] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In **Proceedings of the 34th International Conference on Neural Information Processing Systems**, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[6] Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran T. Tchrakian, Radu Marinescu, Elizabeth M. Daly, Inkit Padhi, and Prasanna Sattigeri. Wikicontradict: A benchmark for evaluating LLMs on real-world knowledge conflicts from wikipedia. In **The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track**, 2024.

[7] Jie Ouyang, Tingyue Pan, Mingyue Cheng, Ruiran Yan, Yucong Luo, Jiaying Lin, and Qi Liu. HoH: A dynamic benchmark for evaluating the impact of outdated information on retrieval-augmented generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 6036–6063, Vienna, Austria, July 2025. Association for Computational Linguistics.

[8] Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. Dense X retrieval: What retrieval granularity should we use? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 15159–15177, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[9] Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. Retrieval augmented generation or long-context LLMs? a comprehensive study and hybrid approach. In Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track**, pp. 881–893, Miami, Florida, US, November 2024. Association for Computational Linguistics.

[10] Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng. Context length alone hurts LLM performance despite perfect retrieval. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, **Findings of the Association for Computational Linguistics: EMNLP 2025**, pp. 23281–23298, Suzhou, China, November 2025. Association for Computational Linguistics.

[11] OpenAI. Gpt-4 technical report, 2024.

[12] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. 2021.

[13] Aaron Grattafiori, Abhimanyu Dubey, and Abhinav Jauhri. The llama 3 herd of models, 2024.

[14] Belinda Z. Li, Emmy Liu, Alexis Ross, Abbas Zeitoun, Graham Neubig, and Jacob Andreas. Language modeling with editable external knowledge. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, **Findings of the Association for Computational Linguistics: NAACL 2025**, pp. 3070–3090, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[15] Zhiyuan Zhu, Yusheng Liao, Zhe Chen, Yuhao Wang, Yunfeng Guan, Yanfeng Wang, and Yu Wang. EvolveBench: A comprehensive benchmark for assessing temporal awareness in LLMs on evolving knowledge. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 16173–16188, Vienna, Austria, July 2025. Association for Computational Linguistics.

```
[P122]
Date: 2021-09-16
Statement: Anner-Marie Trevelyan is back
in the Cabinet.
Relations:
- This statement updates the following
earlier propositions: P104 (2021-09-01),
P107 (2021-09-11)
- This statement is updated by the
following later propositions: P151
(2021-10-01), P593 (2022-09-01)
- This statement is supported by the
following later propositions: P513
(2022-09-01)
```

図 3: 命題と命題間関係を，テキストコンテキストとして言語化する例．図中の内容は，実際に抽出された命題および関係に基づく．

A　手法に関する補足

本研究では，命題および命題間関係を，QA モデルが直接利用可能なテキストコンテキストとして言語化する．図 3 は，タイムスタンプ付き命題とその更新・矛盾・根拠関係が，どのような形式でテキストとして表現されるかを示す．

B　データセット

Asteria は，時間とともに変化し未整理な外部知識に対する推論能力を評価するために構築した時間的質問応答データセットである．Asteria では，時間的更新への対応，矛盾の解消，および誤情報の排除が要求される．LLM の内部知識への依存を避けるため，完全に架空の世界を設計し，その内部で外部知識ベースを生成する．具体的には，Gemini 3 Flash を用いて，エネルギー危機，新規疾患の流行，政治組織内の汚職など 20 の主要イベントからなる架空世界を構築する．各イベントについて，事象の進展更新，相互に矛盾する主張，および偽情報を含む 20 個のタイムスタンプ付き命題からなる時系列を作成する．次に，各イベントに対して 3 つの質問文を生成し，それぞれに対して時間的に異なる 3 つの正解を対応付ける．質問に正しく答えるためには，誤情報を排除し，根拠に基づいて矛盾を解消し，事実の時間的変化を追跡する必要がある．Gemini 3 による検証と人手確認を経た後，各質問を多肢選択形式の QA として表現する．同一の質問文に対して，時間枠に応じて正解が変化するため，選択肢はいずれかの時間における正解を含む．最終的に，Asteria は 180 件のユニークな時間的質問応答ペアから構成される．

また，より現実的な設定でも評価を行うため，Web 検索に基づくニュース記事から構築された実世界の時間的 QA ベンチマークである **CLARK-News** [14] を用いる．CLARK-News では，前処理で，同一質問文で正解が時間とともに少なくとも 3 回以上変化する事例のみを残し，同一時間において複数の正解を持つ質問を除外する．そして，すべての質問は Asteria と同一の手順により多肢選択形式へ変換する．前処理後，CLARK-News は 193 件のユニークな時間的質問応答ペアから構成された．