物語性は大規模言語モデルの長文脈記憶を安定化させるか？

大谷 紗慧 [1] 村山 太一 [1]
[1] 横浜国立大学 理工学部
otani-sae-bg@ynu.jp
murayama-taichi-bs@ynu.ac.jp

概要

大規模言語モデル（LLM）は長文脈からの情報想起が不安定である一方，人間は出来事を構造化して記憶することで頑健な想起を行う．本研究では，出来事を一貫したエピソードとして組織化する性質である物語性に注目し，その程度が LLM の記憶の頑健性に与える影響を検証する．我々は，イベント構造を制御した高物語性・低物語性文章からなるデータセットを構築し，複数の干渉条件下で想起タスクを評価した．その結果，物語性は想起精度を一様に向上させないが，意味的に競合する干渉の下での性能劣化を緩和することが明らかになった．さらに内部表現の分析から，高物語性文脈において Attention が表層的手がかりへの過剰集中を避け，出来事間の関係構造に基づいた情報アクセスが維持されることが示唆された．本研究は，談話レベルの構造が長文脈下における LLM の記憶挙動に重要な役割を果たすことを示す．

1 はじめに

大規模言語モデル（LLM）は，対話，検索，要約など多くのユーザ向けアプリケーションの基盤技術となっており，長い入力文脈や対話履歴を統合した応答生成が求められる．しかし先行研究により，LLM は入力文脈が長くなるにつれて重要な情報の保持・想起が困難となり，意味的に重要だが目立たない情報を見落としやすいことが示されている．このような記憶の不安定性は，下流タスクにおける一貫性の低下を引き起こす要因となる．長文脈処理におけるこのボトルネックを克服するためには，LLM の記憶の頑健性を規定する要因を明らかにする必要がある．

LLM の記憶特性に関する研究は，主に位置バイアスや注意減衰，文脈長感度といった構造的要因に焦点を当ててきた [1, 2, 3, 4]．これらは，情報の意味内容よりも，提示される位置や形式が判断や想起に強く影響することを示している．一方で，文脈全体の談話構造が想起に与える影響については，十分に検討されていない．この点で手がかりとなるのが，人間の記憶研究で重視されてきた物語性である [5]．物語性は，出来事同士を結び付け，一貫したエピソードとして構造化する性質であり，このような組織化は人間の想起過程において有効に機能する．この知見は，入力文脈を物語的に組織化することが，LLM においても長文脈処理の頑健化に資する可能性を示唆している．

本研究では，文章の物語性の程度が LLM の長文脈記憶を安定化させるという仮説を検証する．具体的には，高物語性および低物語性の文章からなるデータセットを構築し，複数の干渉条件下において想起タスクを実施することで，LLM の行動レベルでの記憶性能および内部表現の程度を評価する．これらの取り組みを通じて，本稿では以下の 3 つの研究課題に取り組む：

RQ1: 物語性の違いは，LLM の想起性能に差をもたらすか？ (§4.1)

RQ2: 時間，場所，人物，内容といった要素のうち，どれが物語性の影響を受けやすいか？ (§4.2)

RQ3: 物語性は，LLM の内部表現をどのように形成・変化させるか？ (§5)

これらの問いを通じて，本研究は，入力文脈における出来事の談話的組織化が，LLM の記憶の頑健性に果たす役割を明らかにすることを目指す．

2 背景

LLM における長文脈記憶 LLM の文脈長は近年大きく拡張されているが，依然として長い入力文脈に含まれる情報を安定して保持・活用することは困難である．長文脈処理能力の評価は，無関係な情報の中に重要情報を埋め込む Needle-in-a-

Haystack（NIAH）パラダイムを中心に発展してきた [6, 7, 8, 9]．一方，近年は単なる長文脈中に埋め込まれた情報の想起性能だけでなく，長文脈内の情報がどのように参照されるかという想起過程そのものに注目が移っている．その結果，注意配分が入力位置に依存する系統的なバイアスを示し，特に文脈中央部の情報や妨害文の存在下で想起が不安定になることが報告されている [1, 2]．さらに，文脈長や構造的複雑性の増大に伴って性能低下が進行する現象は **context rot** と呼ばれ [3, 4, 10, 11, 12]，長文脈処理が一様な記憶空間ではなく，構造的・位置的制約に強く影響されることを示している．

人間認知と LLM 表現の橋渡し　人間の認知研究は，出来事を因果的・時間的に組織化した物語構造が，記憶の形成と検索を支える重要な談話レベルの要因であることを示している [5, 13, 14, 15, 16, 17]．大規模テキストで学習された LLM は，構文的・意味的構造に加え，時間的・因果的関係といった高次の関係構造を内部表現として獲得していることが示されている [18, 19, 20]．さらに，神経科学および認知モデリングの研究は，LLM の内部表現が談話レベルや物語理解において人間の認知過程と部分的な類似性を示すことを報告している [21, 22, 23]．これらの知見は，物語性のような組織化原理が，LLM においても安定した表現形成や記憶アクセスに関与する可能性を示唆する．本研究はこの視点に基づき，物語性が妨害下の長文脈処理において，LLM の記憶アクセスを支える組織化の足場として機能するかを検証する．

3　データセット構築と実験設定

本研究の目的は，長文脈入力に対して物語的構造を付与することが，LLM の想起性能を改善するかを検証することである．図 1 は，本研究で用いる実験パイプラインの概要を示す．本研究では，イベント内容を一定に保ったまま物語性（高／低）を操作した文章データセットを構築し，干渉条件下での想起性能を評価する．

3.1　データセット構築

本研究では，物語性を「出来事が，時間的・因果的関係および主体の連続性に基づいて一貫した構造を形成する度合い」と定義する．人間のエピソード記憶に着想を得て，各章は時間順に並んだ離散的なイベント列として表現される [24]．各イベント e_i

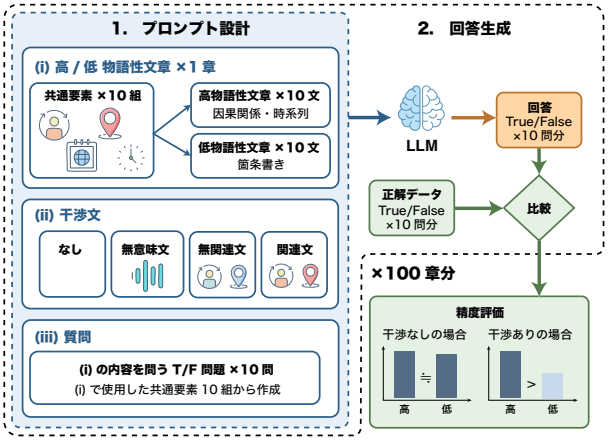


図 1: 実験設定の概要．本実験は，LLM の長文記憶において，物語性が記憶の安定性を高めるかどうかを検証する．同一のイベント集合に対して高／低物語性の文章を構築し（1），干渉条件を付加した上で真偽値質問に回答させる（2）．これを 100 章分にわたり，評価する．

は，時間情報 t_i，場所情報 s_i，関与主体 ent_i，および内容 c_i からなるタプルとして定義される：

$$e_i = (t_i, s_i, ent_i, c_i). \tag{1}$$

データセットの基本単位は 1 章であり，各章は 10 個のイベントから構成される．各イベント要素は，事前定義された語彙集合からテンプレートに基づいて割り当てられる．1 つの章に対して，高物語性文章と低物語性文章の 2 種類を生成する．高物語性条件では，時間順序，因果関係，および主体の連続性を保持した一貫した叙述を行う一方，低物語性条件では，因果関係を除去し，イベント順序をシャッフルすることで断片的な構造を作成する．

各章に対して，時間・場所・主体・内容の各要素に対応する 10 個の真偽値質問を付与する．本研究では M=100 章のエピソードを用い，物語性（高／低）と干渉条件（NI/MI/UI/RI）の全組合せに対して評価を行う．したがって，総質問数は $M \times 2 \times 4 \times 10$ となる．

3.2　実験設定

評価時には，LLM に対して (i) 高または低物語性の文章，(ii) 条件に応じた干渉文，(iii) 各イベント要素を対象とする 10 個の真偽値質問を入力する．干渉条件として，以下の 4 種類を設定する．**NI**（No Interference）は干渉文を付加しない条件である．**MI**（Meaningless Interference）は，意味を持たない単語群

Cond.	Narrativity	Acc	Rec	Prec	F1
NI	High	0.546	0.546	0.548	0.542
	Low	0.565	0.565	0.566	0.564
MI	High	0.537	0.537	0.538	0.532
	Low	0.518	0.518	0.518	0.516
UI	High	0.538	0.538	0.540	0.534
	Low	0.543	0.543	0.544	0.540
RI	High	0.546	0.546	0.548	0.541
	Low	0.530	0.530	0.530	0.528

表 1: 各物語性と各干渉条件を組み合わせた精度比較．各条件ごとに 100 章分集計した結果を報告．Recall，Precision，F1 スコアは全問題にわたるマクロ平均値．

を挿入する条件である．**UI**（Unrelated Interference）は，対応する章と無関係な内容をもつ自然言語文を挿入する条件である．**RI**（Related Interference）は，対応する章とイベント構成要素の一部を共有するが異なる出来事を記述した文を挿入する条件である．モデルの出力は真偽値応答として取得し，条件ごとに想起精度を算出する．

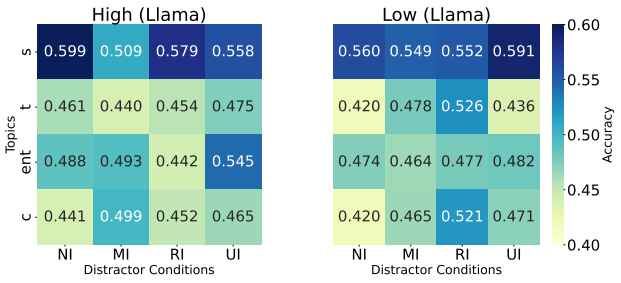
断りのない限り，本実験では **Llama-2-13b-chat-hf** を用いる．結果は全章にわたって集約し，物語性および干渉条件間で比較することで，長文脈下における記憶の頑健性を評価する．

4 結果

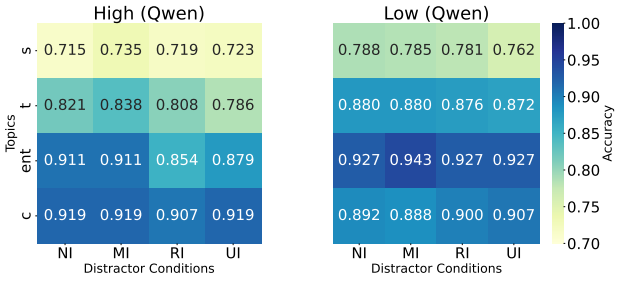
本節では，RQ1 および RQ2 に対応する実験結果を報告する．まず，物語性の違いが想起性能に与える影響を検討し（4.1 節），次に，どのイベント要素がこれらの差異に寄与しているかを要素別に明らかにする（4.2 節）．

4.1 物語性による想起性能の差異

表 1 は，Llama-2-13b-chat-hf の物語性および干渉条件ごとの全体的な想起性能を示す．結果から，物語性の影響は一様ではなく，干渉の有無および種類に依存することが分かる．まず，NI 条件では，高物語性・低物語性間の性能差は小さく，指標間での差異も限定的である．同様に，UI 条件においても，物語性による性能差は小さく，両条件でほぼ同程度の成績が観察された．一方，MI および RI 条件では，高物語性文章が低物語性文章を一貫して上回る傾向



(a) Llama-2-13b-chat-hf



(b) Qwen2.5-14B-Instruct

図 2: 異なる干渉条件下における要素ごとの想起精度．各モデルについて，高物語性文章と低物語性文章それぞれに対し，イベント構成要素および干渉条件別に想起精度を示す．

が確認された．特に RI 条件では，高物語性条件において想起率および F1 スコアが顕著に向上しており，意味的競合が生じる状況下において，物語性が記憶の頑健性を効果的に支えることが示唆される．これらの結果は，物語性が記憶性能を一様に向上させるのではなく，意味的競合が生じる状況下において記憶の頑健性を高めることを示している．

4.2 要素別の物語性効果

図 2 は，時間，場所，主体，内容の各要素に対する想起精度を，モデルおよび干渉条件別に示す．全体として，物語性の効果は要素間で均一ではなく，モデル依存的であることが確認された．Llama-2-13b-chat-hf では，場所情報に関する想起が高物語性条件下で安定して向上する一方，他の要素では差異は限定的であった．一方，Qwen2.5-14B-Instruct では，物語性の影響を受ける要素の傾向が異なり，主体および内容情報に対する性能は比較的安定していた．これらの結果は，物語性が特定のイベント要素に対する内部表現や検索挙動を選択的に安定化させること，およびその影響がモデルごとに異なることを示唆している．

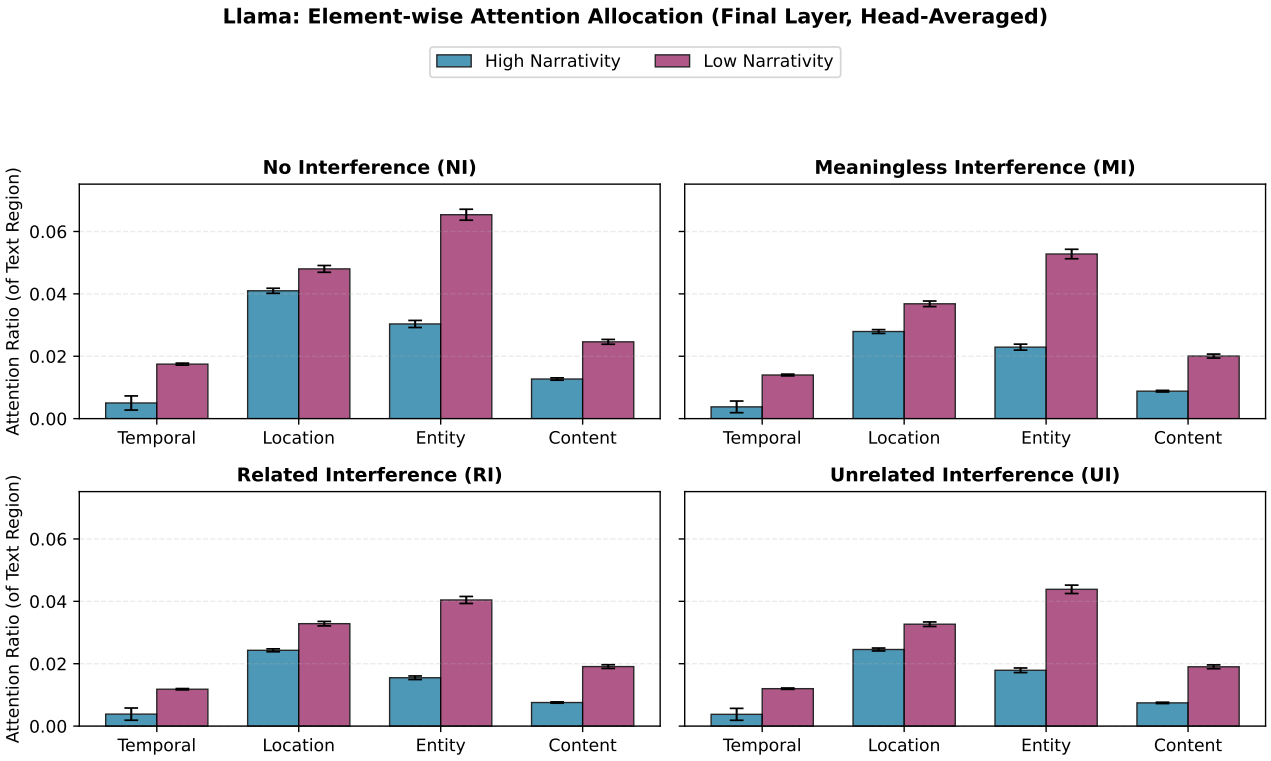


図 3: Llama-2-13b-chat-hf の最終 Transformer 層における，最終クエリ位置から入力トークンへの Attention 分布．Attention 重みはヘッド間で平均化し，章間で集約している．異なる干渉条件下における，各条件について，時間，場所，主体，内容のカテゴリ別に結果を示す．

5　内部記憶構造の表現レベル分析

本節では，RQ3 に対応して，質問応答時に LLM が入力文脈のどの情報に相対的に注意を向けているかを調べるため，最終層の注意分布を分析する．具体的には，質問トークンから入力トークンへの Attention 重みを抽出し，時間，場所，主体，内容，干渉文，その他の 6 カテゴリに分類して集約する．これにより，文脈長やトークン数の違いを超えて比較可能な Attention 分布プロファイルを得る．

図 3 に示す通り，Attention の割り当ては物語性および干渉条件に応じて体系的に変化する．低物語性条件では，特に RI 条件下において，主体（entity）や場所（location）といった局所的・表層的な手がかりに対する Attention が顕著に増加する傾向が見られる．一方，高物語性条件では，Attention は特定の要素に過度に集中することなく，時間・場所・主体・内容といった複数の要素に比較的均等に分散する．NI 条件および UI 条件では，物語性間の Attention 分布の差は比較的小さく，4.1 節で観察された性能差の小ささと対応している．

これらの結果は，物語性が Attention の総量を単純に増加させるのではなく，どの情報要素に注意を向けるかという検索の配分構造を調整することで，特に意味的競合が生じる干渉下において，文章全体に基づく安定した記憶アクセスを支えていることを示唆している．

6　おわりに

本研究では，人間の記憶において安定性を支える組織化原理である物語性が，長文脈条件下における LLM による想起に与える影響を検証した．イベント構造を明示的に制御した物語性文章データセットを構築し，複数の干渉条件下で想起性能および内部表現を評価した．その結果，物語性は記憶性能を一様に向上させるわけではなく，干渉の有無および性質に依存して効果を示すことが明らかになった．特に，意味的に競合する干渉が存在する場合に，高物語性文章は想起の頑健性を一貫して向上させた．さらに表現レベルの分析から，物語性は Attention 分布や検索挙動を通じて，高物語性情報へのアクセス構造を安定化させることが示唆された．

謝辞

本研究は JSPS 科研費 JP23K16889 の助成を受けたものです.

参考文献

[1]Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the Emergence of Position Bias in Transformers. In **Forty-Second International Conference on Machine Learning**, June 2025.

[2]Zihao Yi, Delong Zeng, Zhenqing Ling, Haohao Luo, Zhe Xu, Wei Liu, Jian Luan, Wanxia Cao, and Ying Shen. Attention Basin: Why Contextual Position Matters in Large Language Models, August 2025.

[3]Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. **Transactions of the Association for Computational Linguistics**, 12:157–173, 2024.

[4]Runchu Tian, Yanghao Li, Yuepeng Fu, Siyang Deng, Qinyu Luo, Cheng Qian, Shuo Wang, Xin Cong, Zhong Zhang, Yesai Wu, Yankai Lin, Huadong Wang, and Xiaojiang Liu. Distance between Relevant Information Pieces Causes Bias in Long-Context LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Findings of the Association for Computational Linguistics: ACL 2025**, pages 521–533, Vienna, Austria, July 2025. Association for Computational Linguistics.

[5]Jean M. Mandler. **Stories, Scripts, and Scenes: Aspects of Schema Theory**. Psychology Press, London and New York, 2014. First published 1984; expansion of lectures on schema theory.

[6]Elliot Nelson, Georgios Kollias, Payel Das, Subhajit Chaudhury, and Soham Dan. Needle in the Haystack for Memory Based Large Language Models, July 2024.

[7]Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. Multimodal Needle in a Haystack: Benchmarking Long-Context Capability of Multimodal Large Language Models, February 2025.

[8]Hyeonseok Moon and Heuiseok Lim. NeedleChain: Measuring Intact Long-Context Reasoning Capability of Large Language Models, July 2025.

[9]Yunfan Gao, Yun Xiong, Wenlong Wu, Zijing Huang, Bohan Li, and Haofen Wang. U-NIAH: Unified RAG and LLM Evaluation for Long Context Needle-In-A-Haystack, March 2025.

[10]Chroma. Context rot: How increasing input tokens impacts llm performance. https://research.trychroma.com/context-rot, 2025. Accessed: 2025-12-09.

[11]Reduce Context Rot in LLMs with RAG & DeepSearch | Valyu. https://www.valyu.network, August 2025.

[12]Understanding and Solving Context Rot in LLMs: Why Longer Inputs Often Fail, and What You Can Do About It - wissly | Custom AI Document Solution for Learning All Your Documents. https://www.wissly.ai/en/blog/context-rot-in-llms-and-how-to-fix-it.

[13]Gordon H. Bower and Michal C. Clark. Narrative stories as mediators for serial learning. **Psychonomic Science**, 14(4):181–182, April 1969.

[14]Mark J. Huff and Glen E. Bodner. All varieties of encoding variability are not created equal: Separating variable processing from variable tasks. **Journal of memory and language**, 73:43–58, May 2014.

[15]Michelangelo Naim, Mikhail Katkov, Stefano Recanatesi, and Misha Tsodyks. Emergence of hierarchical organization in memory for random material. **Scientific Reports**, 9(1):10448, July 2019.

[16]J. M. Mandler and N. S. Johnson. Remembrance of things parsed: Story structure and recall. **Cognitive Psychology**, 9:111–151, 1977.

[17]T. Trabasso and L. L. Sperry. Causal relatedness and importance of story events. **Journal of Memory and Language**, 24:595–611, 1985.

[18]Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT Rediscovers the Classical NLP Pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pages 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics.

[19]John Hewitt and Christopher D. Manning. A Structural Probe for Finding Syntax in Word Representations. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pages 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[20]Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language Models as Knowledge Bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pages 2463–2473, Hong Kong, China, January 2019. Association for Computational Linguistics.

[21]Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?, January 2020.

[22]Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. Long-range and hierarchical language predictions in brains and algorithms, November 2021.

[23]Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. **Communications Biology**, 5(1):134, February 2022.

[24]Zafeirios Fountas, Martin Benfeghoul, Adnan Oomerjee, Fenia Christopoulou, Gerasimos Lampouras, Haitham Bou Ammar, and Jun Wang. Human-inspired episodic memory for infinite context LLMs. In **The Thirteenth International Conference on Learning Representations**, 2025.

A　データセットについて

A.1　物語性操作の具体例

イベント要素	内容
Location	Throggs Neck
Temporal	2017-08-05
Entity	Gabriella Scott
Content	Jazz festival in Harlem
高物語性文章	Gabriella Scott, on the fifth day of August in the year 2017, found herself enraptured at the heart of Throggs Neck, where the melodies of the Gabriella Jazz festival in Harlem danced upon the air, carrying souls to heights unseen.
低物語性文章	At Throggs Neck on 2017-08-05, Gabriella Scott visited the Jazz festival in Harlem.

表 2: 共通要素によって表現される高物語性文章と低物語性文章の例.

表 2 は，共通のイベント要素集合から生成された高物語性及び低物語性文章の例である．イベント要素及び高物語性文章の生成には `gpt-3.5-turbo` を使用した．

A.2　干渉文の例

関連干渉（RI）では，元のエピソードと主体を共有するが，異なる出来事を記述する文を挿入する．

> Gabriella Scott was in Sunset yoga in Battery Park at Whitney Museum of American Art on 2017-05-29. ...

A.3　質問例

各エピソードは 10 個のイベントから構成される．以下に，1 イベントに対応する True/False 質問例を示す．

> Is it true that what Gabriella was occured at Throggs Neck on 2017-08-05 is Jazz festival in Harlem.?
> → True

B　他モデルでの検証

本節では，Llama と同様の手法を用いて，Qwen2.5-14B-Instruct における最終層 Attention 分布を分析し，物語性および干渉条件が内部表現に与える影響を検討する（図 4）．

Qwen は Llama とは質的に異なる Attention パターンを示した．すべての干渉条件において，低物語性文章は時間情報への Attention が一貫して高く，高物語性文章では時間情報への注意は相対的に低かった．この傾向は干渉の有無や種類に依存せず，安定して観察された．一方，場所情報に対する Attention は逆の傾向を示し，高物語性文章で一貫して高かった．主体および内容に対する Attention についても類似した傾向が見られたが，差は比較的小さかった．物語性による差異は，干渉条件に依存した動的な検索戦略というよりも，モデル固有の比較的静的な Attention 配分の違いとして現れている．最終層 Attention は記憶アクセスの一側面を捉える有用な指標ではあるが，その振る舞いが想起性能にどのように結び付くかはモデルごとに異なることが示唆される．

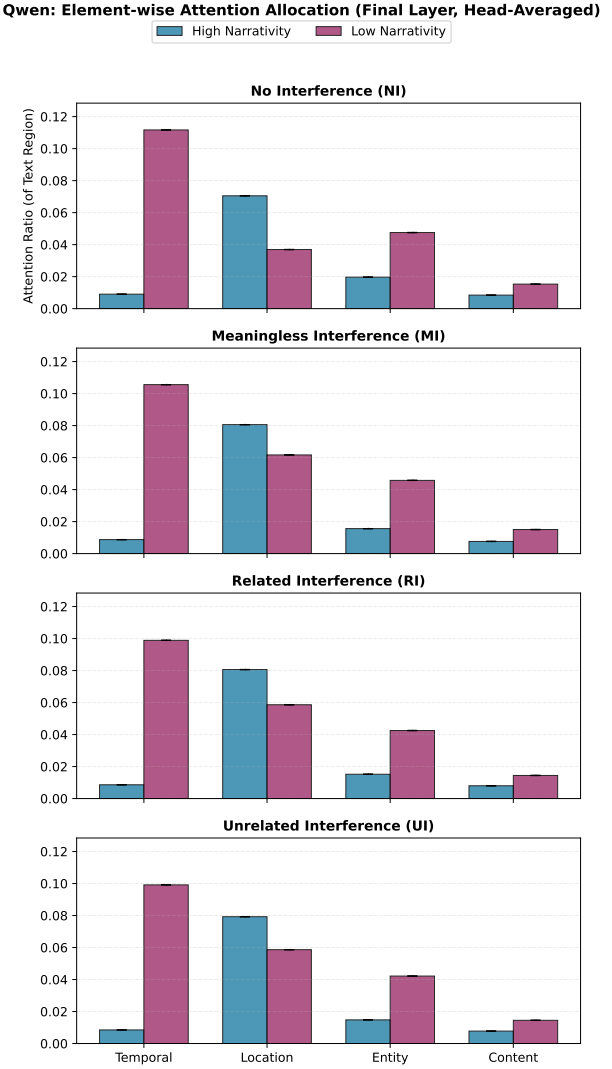


図 4: Qwen2.5-14B-Instruct の最終 Transformer 層における，最終クエリ位置から入力トークンへの Attention 分布.