プロンプトへのフレーム知識付与による推論タスクの性能向上

岩本蘭 [1,2]　小原京子 [1]
[1] 慶應義塾大学　[2] 日本アイ・ビー・エム株式会社
r.iwamoto@keio.jp　ohara@hc.st.keio.ac.jp

概要

認知や身体性に根ざした，人間の持つ言葉の意味に関する知識を大規模言語モデルに組み込む研究は未だほぼ皆無である．本研究では，人間の持つ，経験と結びついた意味に関する知識であるフレーム知識をプロンプトとして与えることで，推論タスクの性能向上をはかる．映画の一場面のキャプションから次の場面のキャプションを選択する推論タスクに対し，付与するフレーム知識の種類と範囲を変化させ，アウトプットの精度が向上するかを調べる実験を行った．その結果，フレーム名とフレーム要素を併記した条件で精度が最も向上した．

1　はじめに

現在の大規模言語モデル (LLM) は，人間とは異なり身体的感覚や経験と結びついた知識を持っているわけではない [1]．その結果，共起や頻度に基づきアウトプットは生成できるが，出来事の背後にある意味を扱うことが難しい．FrameNet [2, 3] は人間の経験に根ざした，人間の持つ言葉の意味に関する知識を記述した資源である．フレーム意味論 [4, 5] に基づいて，言葉の意味をフレーム知識 (様々な出来事や行為に関する背景知識) として記述する．例えば英語の動詞 *buy* と *sell* はいずれも `Commercial_transaction` フレーム（買い手と売り手が商品を貨幣と交換することに関する背景知識）を喚起すると捉える．現在の LLM はフレーム知識を部分的に内包している可能性が指摘されている [6] が，タスクにどのようなフレーム知識が有効かはまだ明らかになってはない．

本研究は，フレーム意味論に基づくフレーム知識（フレーム名/フレーム要素/定義文）をプロンプトに埋め込み，LLM へのフレーム知識付与の有効性を検証する．具体的には，出来事の連続性や因果関係を判断する推論課題に対して，プロンプト内のフレーム知識の種類（フレーム名/要素/定義文）と付与範囲（主述語のみ/全内容語）を様々に変化させ，どの条件が出力の精度向上に貢献するかを調べた．

本論文の貢献は以下の通りである．

- フレーム知識をプロンプトに付与し，推論タスクでの有効性と，アノテーションの種類や範囲によって性能が変化することを示した．
- 構文情報や WordNet といった他の言語知識を付与する場合と比較して，フレーム知識の付与が高い推論性能をもたらすことを示した．

2　提案手法

LLM のプロンプトにフレーム知識，構文情報，WordNet [7, 8] 情報を付与し，推論性能を比較した．

2.1　対象タスク: SWAG

本研究では SWAG (Situations With Adversarial Generations) [9] タスクを用いた．これは実際の映画のキャプションを前提文として与え，次の場面のキャプションを 4 択で選ぶ，状況理解や因果関係に基づく推論が要求されるタスクである．誤答候補は自動生成とフィルタリングを組み合わせた adversarial な手法で作成されている．以下に問題例を示す．

前提文: *That player then throws the ball across to a man on the left.*
1: *A man throws a ball on the mat with other people watching in the background.*
2: *A man kicks a ball into his crotch.*
3: *A man grabs the ball and throws it in the goal.*
4: *A man in white shirt throws a ball around another boy while the man throws another ball.*

2.2　実験準備

frame-semantic-transformer [10] を用いてプロンプトにフレーム知識を付与した．T5 [11] ベースの本解析器は文中の語を同定し，対応するフレームとフレーム要素（Frame Element; FE）を推定する．フ

レーム分類の精度は 89%, フレーム要素抽出の精度は 75%である．本研究ではフレーム名・FE およびフレーム定義文の最初の 3 文を付与した．

2.3 実験設定

本実験では，SWAG 検証データ 5000 問から低品質な文を含む問題や多数の選択肢に同じフレームがつく問題を除外した[1]．残った 1879 問にアノテーション (注釈) としてフレーム情報 (提案手法)，構文情報，WordNet 情報の 3 種類を付与した．各言語情報の注釈は候補文に付与し，(i) 全内容語に付与，(ii) 構文解析により同定した依存構造の Head (本論文では主述語と呼ぶ) のみに付与する 2 種類の条件を設けた．(ii) の注釈例を下に示す．

候補文の例: *The man lifts the barbell up to his chest.*

提案手法: フレーム

- **フレーム名** `Cause_motion` [LU: *lifts*]
- **フレーム名&FE** `Cause_motion` [LU: *lifts*, FE: AGENT: *The man*; THEME: *the barbell*; GOAL: *up to his chest*.]
- **フレーム名&定義** `Cause_motion` [LU: *lifts*, Definition: An Agent causes a Theme to move from a Source, along a Path, to a Goal.]

比較手法: 構文 *lift* [nsubj: *man*, obj: *barbell*] (依存構造に基づくラベル)

比較手法: WordNet *lift* [raise from a lower to a higher position] (synset および定義)

構文注釈では stanza [12] を用いて述語を中心とした依存関係を取得した．WordNet 情報を付与する条件では名詞，動詞，形容詞，副詞を対象とし，語義曖昧性解消のために Lesk 法 [13] を用いて，定義文・要例文と候補文との単語重なりが最大となる synset を選んだ．本研究では推論挙動の安定性を重視し，API で温度制御可能な GPT-4o-mini[14] を用いた．プロンプトでは前提文と候補文を提示した後に候補文の注釈を付与した (付録 A.1 を参照)．正答率を指標とし，選択肢の提示順をランダムに置換する乱数シードを 5 通り変更して実行した平均を報告する．

3 結果

本節ではフレーム知識の付与が LLM の推論性能に与えた影響を分析する．まずアノテーションの種類/付与範囲の違いが推論精度に及ぼす影響を検討し，次に FrameNet と他の言語知識資源を比較する．

1) 詳細は付録 A.2 を参照

表 1 SWAG におけるフレーム知識の付与条件の比較

条件（FrameNet）	付与範囲	正答率
なし（Baseline）	–	75.2 ± 0.3
フレーム名	全内容語	79.8 ± 0.3
	主述語のみ	79.9 ± 0.4
フレーム名&FE	全内容語	79.5 ± 0.5
	主述語のみ	**80.3** ± 0.6
フレーム名&定義	全内容語	79.5 ± 0.6
	主述語のみ	79.9 ± 0.7

3.1 フレーム情報付与条件の比較

フレーム情報の種類（フレーム名，FE，定義）と付与範囲（全内容語/主述語のみ）を変化させた比較結果を表 1 に示す．いずれの条件でも Baseline (75.2%) と比較して一貫して性能が向上した．付与範囲による性能差は小さいものの，多くの条件において主述語のみに付与した方がわずかに高い正答率を示した．また，フレーム名に FE を追加した場合には主述語のみ付与条件で性能がさらに向上し，80.3% と最も高い正答率が得られた．

3.2 言語知識資源の比較

次にフレーム知識と他の言語知識，構文情報と WordNet 情報を付与した場合の性能比較を表 2 に示す．フレーム情報については前節で用いた条件のうち最も性能が高かった，フレーム名とフレーム要素を主述語のみに付与する設定を採用した．以降の実験でも同じフレーム情報条件を用いる．いずれの言語情報を付与した場合でも Baseline と比較して性能が向上したが，フレーム情報を付与した条件では構文情報や WordNet と比較して改善幅が大きかった．構文情報の効果が限定的であった理由として，LLM は主語・目的語といった基本的な構文関係を比較的正確に扱えることが報告されており [15]，追加情報として十分に機能しなかった可能性がある．また WordNet 上の情報は語義の区別や整理には有効である一方，出来事の因果関係や時間的な連続性を直接表現しない．そのため，SWAG のような次の行為を推論するタスクには有効ではないと考えられる．これに対しフレーム知識は，出来事のタイプや役割と

表 2　SWAG における言語資源の効果比較

言語知識資源	正答率
なし（Baseline）	75.2 ± 0.3
構文	76.2 ± 0.3
WordNet	77.3 ± 0.5
フレーム名&FE	**80.3** ± 0.6
フレーム名&構文	79.9 ± 0.6
フレーム名&FE&構文	79.8 ± 0.6
フレーム名&定義&構文	79.4 ± 0.7

いった事象構造を明示的に表現するので，本タスクでの推論に直接寄与した可能性がある．フレーム情報と構文情報を併用した場合は性能向上は確認されず，正答率はわずかに低下した．

4　分析

4.1　フレーム知識付与による正答例

本節では，フレーム知識付与による回答改善を定性的に分析する．対象として，3.1 節で最高性能を示した，選択肢の主述語のみにフレーム名とフレーム要素を付与する条件を用いる．なお，本来は 4 択だが，紙面の都合上，Baseline が選択した誤り文とフレーム条件が選択した正解文のみを示す．

以下の 3 つの例から，フレーム知識はモデルの判断基準を語彙一致や共起という表層表現から因果関係を考慮した基準へと変化させると考えられる．

4.1.1　予想される後続状況の事象タイプ

この設問の前提文ではプロペラの減速という異常な状態変化が提示されており，次の出来事としてはそれへの対処行為が想定される．

前提文:*Underwater, the ship's main propeller slows.*
誤り文:*An engineer cautiously approaches the cabin panel.*
正解文:*An engineer pulls down a thick lever.*

正答文の主述語 *pull* は `Manipulation` フレーム，誤答文の主述語 *approach* は `Arriving` フレームを喚起する．ここでは前提文の状況に直接対処する `Manipulation` フレームの方が適切である．

4.1.2　主述語フレーム間の近接性

前提文の主述語である *raises* と正解文の主述語 *lifts* はどちらも *glass* を FE に持ち，同じフレーム `Cause_change_of_position_on_a_scale` を喚起するのに対し，Baseline が選んだ誤り文の主述語の *smile* が喚起する `Making_faces` フレームは *raises* が喚起するフレームと意味的関係を持たない．

前提文:*Someone raises his glass to someone.*
誤り文:*Someone smiles and pushes out of the chair.*
正解文:*Someone lifts his glass, frowns and then sips.*

4.1.3　視点・知覚フレームの利用

以下で前提文は，映画キャプションらしくカメラの動きを描写している．正解文は述語 *shown* が `Cause_to_perceive` フレームを喚起し，前提文と同様にカメラの視点で描写している．対して誤り文の主述語 *dances* は `Self_motion` フレームを喚起する語であり，前提文と視点が異なる．

前提文:*The camera zooms in on the trumpet player.*
誤り文:*A girl dances on the stage playing the drums.*
正解文:*A girl is shown playing piano.*

4.2　フレーム知識付与の利用特性分析

ここではフレーム注釈の量もしくは正確性が推論性能に与える影響を分析する．

4.2.1　注釈量が推論性能に与える影響

まず，フレーム情報を段階的に削減した際の推論性能の変化を分析した．フレーム名&FE（全内容語）条件を対象とし，フレーム名と FE の組を注釈単位とした．各文の注釈単位数を m ，保持率を ratio $\in \{1, 0.75, 0.50, 0.25\}$ とした．(i) 注釈単位をランダムに $\lceil \text{ratio} \times m \rceil$ 個保持する random 条件と (ii) 構文的重要度順[2)]に保持する syntax-based 条件を設けた．どの場合でも少なくとも 1 つの注釈を残した．

図 1 に結果を示す．random 条件では注釈保持率の低下に伴い正答率も低下した．主述語のみを残した場合よりもいずれの保持率でも正答率が低いことから，推論性能の向上には全ての注釈が必要ではないことが示唆される．一方 syntax-based 条件では保持率を低下させても正答率の低下は小さかった．この結果は重要なフレーム注釈の選択が推論性能の向上に寄与することを示唆している．

2)　主述語，`conj` で接続された述語，目的語，斜格要素，主語の順

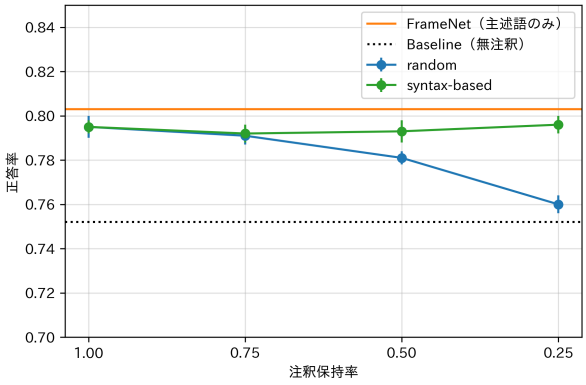

図 1　フレーム注釈保持率が推論性能に与える影響

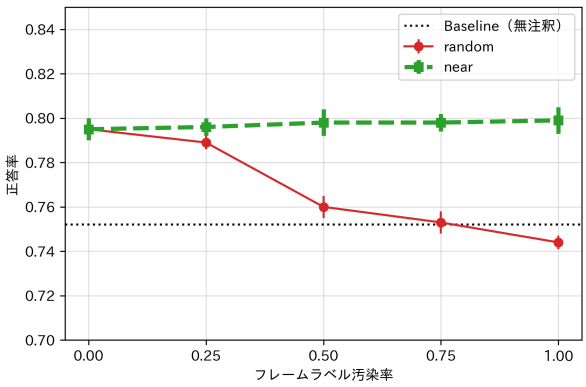

図 2　フレームラベルの正確性が推論性能に与える影響.

4.2.2　フレームラベルの正確性に関する分析

次に，フレーム情報の正確性が推論性能に与える影響を分析した．全内容語に付与した注釈量は固定し，FE と LU を変えずにフレーム名のみを意図的に誤ったラベルに置き換えた．フレーム名の置換には，(i) FrameNet 全体から無作為に選択する random 条件と，(ii) フレーム間関係に基づき元のフレームの近傍フレームを選択する near 条件の 2 種類を用いた．near 条件では，FrameNet 1.7 のフレーム関係から無向グラフを構築し，継承関係や部分フレーム関係で結ばれたフレーム同士を近接フレームとみなした．各注釈単位を確率 $p \in \{0.25, 0.50, 0.75, 1.00\}$ で置換して汚染率を制御し，フレーム名の意味的一貫性が段階的に失われた際の性能変化を分析した．

図 2 は，フレーム名の正確性を段階的に低下させた場合の推論性能を示す．random 条件では汚染率の増加に伴い正答率が単調に低下し，意味的に無関係なフレーム名がノイズになることが確認された．一方 near 条件では汚染率を増加させると正解率はほぼ一定で，全てのフレーム名を置き換えた場合でも Baseline を安定して上回った．この結果は，フレーム名が完全に正確でなくても意味的に近いフレームであれば注釈として有用であることを示している．

表 3　部分正規化信頼度の平均値および中央値

条件	正誤	平均信頼度	中央値
Baseline	正解	0.9749	1.0000
	不正解	0.8724	0.9687
FrameNet	正解	0.9750	1.0000
	不正解	0.8633	0.9562

4.3　推論制御性の分析

フレーム知識の導入がモデルの推論挙動に与える影響を検討するため，正答率に加えて，モデルが各選択肢に割り当てる出力確率に着目した分析を行った．LLM は誤答に対しても高い出力確率を割り当てる過信傾向を示すことが報告されており [16]，精度指標とは独立に，出力確率と正誤結果の関係を分析することが重要である．本研究では，OpenAI API が返す最初の出力トークンに対する対数確率（`logprobs`）を用い，top-k に含まれる選択肢に基づく部分正規化信頼度を算出した．部分正規化信頼度を正誤別・条件別に集計した結果を表 3 に示す．いずれの条件でも正解例では不正解例と比較して高い信頼度が観測されたがその傾向には条件間で差がみられた．Baseline 条件と比較して Frame 条件では正解例での部分正規化信頼度の平均値がわずかに上昇している一方で不正解例における信頼度は低下していた．この結果は，フレーム情報の付与がモデルの判断を一様に強めるのではなく，正解時の確信を補強しつつ，誤答時の過信を抑制する方向に推論挙動を変化させている可能性を示唆している．

5　おわりに

本研究では，人間の持つ，経験や身体性に基づく言葉の意味に関する知識であるフレーム知識が推論タスクにおける LLM の正答率の向上に有効であることを示した．出来事のタイプや役割などを明示的に与えることで LLM の判断基準が表層的な語彙一致から意味的な整合性に基づく推論へと変化する可能性が示唆された．また，フレーム注釈の保持率やラベルの正確性を操作した分析から，推論性能は注釈量ではなく重要度の高い注釈の保持率に依存すると示した．フレーム知識を他の意味知識が必要なタスクで活用することが今後の課題である．

参考文献

[1] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience grounds language. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 8718–8735, 2020.

[2] Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. **FrameNet II: Extended theory and practice (Revised)**. 2016.

[3] Hans C Boas, Josef Ruppenhofer, and Collin F Baker. Framenet at 25: Results and applications. **International Journal of Lexicography**, Vol. 38, No. 2, pp. 159–189, 2025.

[4] Charles J. Fillmore. An Alternative to Checklist Theories of Meaning. In **Proceedings of the First Annual Meeting of the Berkeley Linguistics Society**, pp. 123–131, 1975.

[5] Charles J. Fillmore and Collin Baker. A Frames Approach to Semantic Analysis. In Bernd Heine and Heiko Narrog, editors, **The Oxford Handbook of Linguistic Analysis**, p. 0. Oxford University Press, 2009.

[6] Jayanth Krishna Chundru, Rudrashis Poddar, Jie Cao, and Tianyu Jiang. Do llms encode frame semantics? evidence from frame identification, 2025.

[7] Christiane Fellbaum. A Semantic Network of English: The Mother of All WordNets. **Comput. Humanit.**, Vol. 32, No. 2-3, pp. 209–220, 1998.

[8] George A. Miller. WordNet: A Lexical Database for English. **Communications of the ACM**, Vol. 38, No. 11, pp. 39–41, 1995.

[9] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In **Proceedings of EMNLP**, pp. 93–104, 2018.

[10] David Chanin. Open-source frame semantic parsing. **arXiv preprint arXiv:2303.12788**, 2023.

[11] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[12] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, 2020.

[13] Satanjeev Banerjee and Ted Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In **Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing**, CICLing '02, p. 136–145, Berlin, Heidelberg, 2002. Springer-Verlag.

[14] OpenAI and team. GPT-4 Technical Report, 2024.

[15] Houquan Zhou, Yang Hou, Zhenghua Li, Xuebin Wang, Zhefeng Wang, Xinyu Duan, and Min Zhang. How well do large language models understand syntax? an evaluation by asking natural language questions, 2023.

[16] Fengfei Sun, Ningke Li, Kailong Wang, and Lorenz Goette. Large language models are overconfident and amplify human bias, 2025.

[17] Aaron Grattafiori and teams. The llama 3 herd of models, 2024.

[18] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In **Proceedings of ACL**, pp. 4791–4800, 2019.

A　実験設定

A.1　Prompt

LLM に与えた System prompt と User prompt を示す．

System Prompt

You are solving a multiple-choice continuation task.
You will be given:
- a context sentence
- four possible continuations, each identified only by a number (1, 2, 3, 4)
- optionally, linguistic annotations for Q or each option
Important notes:
- The numbers 1, 2, 3, and 4 have no inherent meaning.
- They do NOT correspond to any conventional ordering or likelihood.
- The correct answer is NOT correlated with the index.
- You must not assume that a smaller index (such as "1") is more likely.
Your task:
1. Understand Q and all options STRICTLY based on meaning.
2. If annotations exist, treat them only as optional hints.
3. The meaning of the original text always has priority.
4. Evaluate all four options fairly.
5. Choose the option that best continues Q.
Restrictions:
- Do NOT choose an answer by pattern or index preference.
- Do NOT randomly guess.
- Do NOT choose based on any ordering or bias.
Your entire output MUST be exactly one character: 1, 2, 3, or 4
No explanation. No reasoning. Only one number. If annotations are provided, treat them only as optional hints about the type of event, and still choose the option that most naturally continues the original text.

User Prompt

Sent: That player then throws the ball across to a man on the left.
1: A man throws a ball on the mat with other people watching in the background.
2: A man kicks a ball into his crotch.
3: A man grabs the ball and throws it in the goal.
4: A man in white shirt throws a ball around another boy while the man throws another ball.
FrameNet frames:
1: cause motion [agent: A man; theme: a ball; goal: on the mat]
2: cause harm [agent: A man; victim: a ball; body part: into his crotch]
3: manipulation [agent: A man; entity: the ball]
4: cause motion [agent: A man in white shirt; theme: a ball; goal: around another boy]

表 4　フレーム注釈付与での正答率のモデル別比較

モデル	Baseline	FrameNet
GPT-4o mini	75.2 ± 0.3	80.3 ± 0.6
GPT-5 mini	80.9 ± 0.4	84.9 ± 0.4
Llama3.3 (70B)	76.3 ± 0.4	79.3 ± 0.3

表 5　HellaSWAG における言語資源の効果比較

知識資源（単独）	正答率
なし（Baseline）	83.2 ± 1.1
構文	83.7 ± 0.4
WordNet	84.7 ± 1.3
FrameNet	85.2 ± 0.9

A.2　評価データのフィルタリング

注釈の効果を測るため，以下の条件で問題を除外した．
低品質な文：語彙多様性が低い文や定動詞がない文，主語と動詞の活用が不整合（例：we/they + VBZ）な文を含む
フレーム重複：3 つ以上の選択肢が同じ主フレームを持つ
フレームなし：主フレーム未付与の候補文が 2 件以上

A.3　アノテーションの前処理

性能向上のため注釈器の出力に以下の前処理を施した．
表記の自然言語化：LU (xxx.v) の品詞を除去し，フレーム名や FE の役割名を小文字化し_を空白に置き換えた．
ノイズフレームの除去：情報量が低い `Adjacency` などのフレームや機能語のフレームは除外した．
Core FE 以外の除去：Peripheral FE を除去した．

B　汎化的挙動の比較

本節では，モデルおよびデータセットの両面からフレーム注釈の一般化性能を検討する．フレーム注釈には，フレーム名と FE を主述語のみに付与する条件を用いた．

B.1　モデル汎化

フレーム注釈による性能向上が GPT-4o mini 以外のモデルでも観測される挙動なのかを調べる．モデルとして GPT-5 mini と Llama3.3(70b) [17] を用いた．結果を表 4 に示す．どのモデルでもフレーム注釈を付与すると性能が向上していることが確認できた．

B.2　データセット汎化

フレーム注釈が他のデータセットにおいても有効かを検証するため，SWAG の派生である HellaSwag [18] を用いた．HellaSwag は SWAG の選択肢の中で表層的な手がかりを元に回答可能なものを除去したより難易度の高いデータセットである．検証データから 900 問を抽出し，SWAG と同一の設定で GPT-5 mini を用いて評価した．

表 5 では，どの言語情報を付与した条件でも Baseline を上回り，特に Frame 情報を付与した条件が最も高い正答率を示した．この結果は，フレーム注釈が表層的手がかりが抑制された設定においても有効であり，出来事の意味的整合性に基づく選択を促す可能性を示唆する．