FrameBench: 意味フレームに基づく意味理解ベンチマーク

矢野千紘　　笹野遼平
名古屋大学大学院情報学研究科
yano.chihiro.j3@s.mail.nagoya-u.ac.jp　sasano@i.nagoya-u.ac.jp

概要

大規模言語モデル (LLM) は高い下流タスク性能を示す一方で、人間と類似した意味解釈を行っているか評価することは難しい。そこで本研究では、LLM の意味理解を評価することを目的とした、フレーム意味論に基づくベンチマーク、FrameBench の構築方法を提案する。FrameBench は同一の動詞が喚起する意味フレームの違いを正しく区別できるかを問う多肢選択式問題で構成され、LLM が表層的な単語の一致だけでなく、人間と類似した意味解釈によって文脈を理解しているかどうか評価する。また、実際に日本語の FrameBench を構築し[1)]、様々なモデルを評価した結果を報告する。小規模なモデルはランダムに近い性能を示す一方で、いくつかの大規模なモデルは人間に迫る性能を達成することが確認された。

1　はじめに

近年、大規模言語モデル (LLM) は急速な発展を遂げており、質問応答や常識推論といった高度な言語理解を要するタスクにおいても高い性能を示している [1]。LLM は人間のように言葉を操る一方で、内部での推論過程はブラックボックスであり、人間と類似した枠組みで意味を解釈している保証はない。LLM の推論過程の不透明性は、一見自然な出力に潜む誤りや矛盾の制御を困難にし、ユーザーとの信頼構築において大きな障壁となる。このような課題を克服するためには、LLM が文章の意味を人間と同じように捉えているかを多角的に評価する必要がある。

本研究では、そのような評価に必要となるベンチマークの 1 つとして、フレーム意味論 [2] に基づき LLM の意味理解を評価するベンチマークである FrameBench を構築する。フレーム意味論とは、言語が表現する出来事や概念をその背景知識である意味フレームと結びつけることで文の意味を分析する枠組みである。LLM の意味理解を評価する既存のベンチマークは下流タスクに基づくものが一般的であり [3, 4, 5]、LLM の評価にフレーム知識を活用する取り組みは限定的である。本研究では特に日本語のフレーム知識である日本語フレームネット [6] を基盤として日本語 FrameBench を構築し、評価を行う。

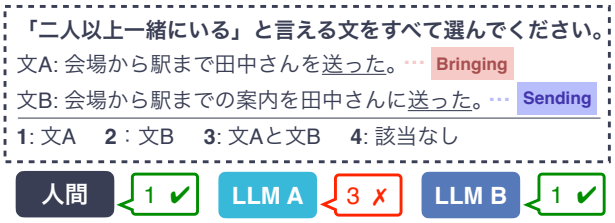

図 1　日本語 FrameBench の例

図 1 に FrameBench に含まれるデータの例を示す。ここでは、動詞「送る」が文脈に応じて異なるフレームを喚起する点に着目している。 例えば、[Bringing] フレームにおいての主体 (Agent) は「対象の移動を制御し、同伴する存在」と定義されるが、[Sending] フレームでの主体 (Sender) は「対象の移動を開始させるが、同伴しない人物」と定義される。FrameBench は、こうした同一の動詞が喚起するフレームの違いを正しく区別できるかを問う多肢選択式問題で構成される。これにより、LLM が表層的な単語の一致だけでなく、人間と類似した意味理解能力を獲得しているかの検証が可能となる。

2　FrameBench

FrameBench はフレーム知識に基づいて、同一の動詞が喚起するフレームの違いを正しく区別できるかを問う問題で構成される。本節では構築手順と、日本語を対象に構築した FrameBench について述べる。

2.1　構築手順

図 2 にベンチマーク構築過程の概要を示す。具体的には以下の 3 手順で構築される。

1) 実装および構築された日本語 FrameBench は https://github.com/SasanoLab/Framebench から利用可能である。

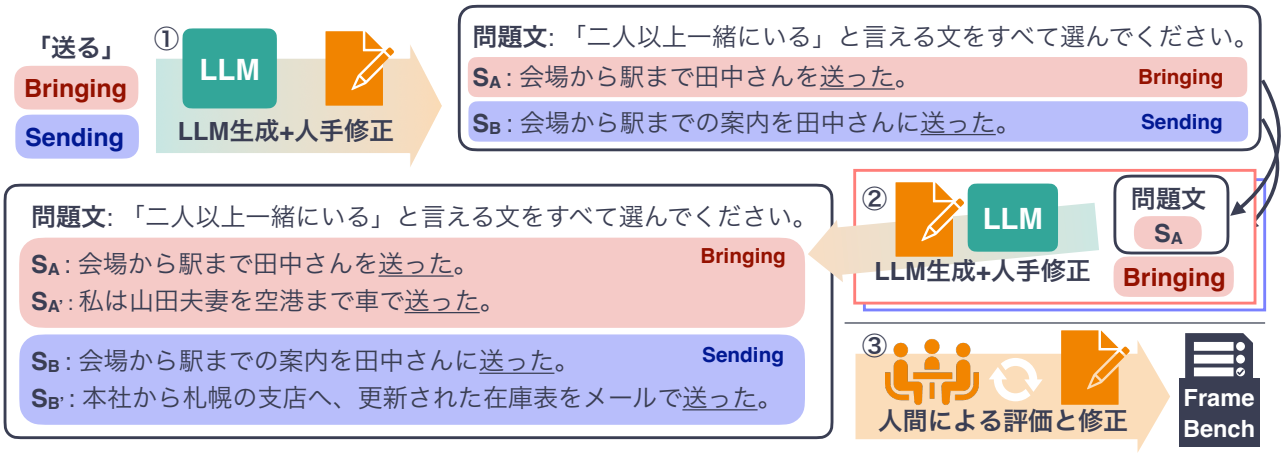

図 2　ベンチマーク構築手順の概要．図中の番号は 2.1 節中の本文に対応している．

手順 ①: 問題文と選択対象文ペアの構築　対象言語のフレーム知識から抽出された複数のフレームを喚起する多義動詞 V と、V によって喚起されるフレームである Frame_A と Frame_B を用いて、問題文と選択対象文ペアを構築する。具体的にはまず LLM を用いて、V を共通の喚起語としながら、Frame_A を喚起する文 S_A と、Frame_B を喚起する文 S_B、および S_A のみが正解となる問題文を生成する。LLM への入力には、ターゲットとなる喚起語 V に加え、2 つのフレームそれぞれの名称、定義、コア要素、用例を含める。また、問題の難易度を担保するため、S_A と S_B は異なるフレームを喚起しつつも、表層的な記述が可能な限り類似するように指示を与える。その後、生成された記述に含まれる不自然な表現の修正を人手で行う。

手順 ②: 選択対象文ペア拡張　表層的に似ていない文ペアを含むことで、評価の多様性を確保できるように、フレームごとに文の追加を行う。具体的には、それぞれのフレームについてフレーム情報と、手順①で生成した問題文および該当のフレームを喚起する文を入力とし、同じフレームを喚起する別の文 $S_{x'}$ を作成する。追加生成された文は入力文と同じフレームを喚起するため、問題文に対する正解および不正解のラベルについても入力文と同じ結果になるように制約を与えて生成する。つまり、Frame_A を喚起する追加の文 $S_{A'}$ を生成する場合は、S_A と同様に問題文に対して正解となるよう制約を与える。生成された文に対しては、手順①と同様に人手による修正を行う。

手順 ③: 人間による評価と修正　構築されたベンチマークの品質を担保するため、対象言語の母語話者による人手評価を実施する。評価内容は、問題文と 2 文に対して正解を「文 1」、「文 2」、「文 1 と文 2」、「該当なし」から選択する 4 択問題への回答と、記述の容認性判定とした。評価対象となる文ペアは、原則として一方が正解で他方が不正解となるペア、つまり (S_A, S_B) もしくは $(S_{A'}, S_{B'})$ のペアを選択するが、必ずどちらか一方が正解であるという推測に基づいた回答を防ぐため、一部のデータにおいて、両方が正解または両方が不正解となるペア、つまり $(S_A, S_{A'})$ もしくは $(S_B, S_{B'})$ のペアを意図的に混入させる。人手評価が完了した後、容認性が低い問題の一部について人手で記述の修正を行い、再度同様の評価プロセスを経てデータセットに採用する。人手評価スコアは各データに 2 組存在する文ペアごとに得られるため、最小値をそのデータの人手評価スコアとする。

最終的に得られた各データは問題文および $S_A, S_B, S_{A'}, S_{B'}$ の 4 文から構成され、正答人数と容認人数からなる人手評価スコアを持つ。

2.2　日本語 Framebench の構築

前節で説明したフレームワークに基づき、日本語 FrameBench データを構築した。フレーム知識として日本語フレームネット [6] を、問題生成及びデータ拡張には GPT-5[2] [7] を利用した。

構築にあたり日本語フレームネットから抽出された、2 つ以上のフレームを喚起する多義動詞は 139 語であり、これらに紐づく利用可能なフレームペアは 335 件[3]であった。手順①において、十分な問題数を確保するため、単一のフレームペアを入力として、2 つの異なる問題を生成した。人手による修正

2)　モデルバージョン: 2025-08-07
3)　例えば 4 つの語義を持つ動詞からは 6 ペアが利用できる

表 1　日本語 FrameBench に含まれる問題の例．太字の文は各問題における正解，下線部はフレーム喚起語，カッコ内は作問時に利用したフレーム名である．これらの情報は人手評価や LLM 評価などで問題を解く際には利用されない．正答モデル数は 3 節と同じモデル群を対象として単一のプロンプトで評価した場合の結果を集計した結果である．

問題	正答モデル数
問題 1.「数値が増えている」と言える文をすべて選んでください。 文 1. シートは加熱ローラーに通され、所定の形に伸ばされた。(Reshaping) **文 2. 販路拡大策が功を奏し、月間受注件数を伸ばした。(Cause_change_of_position_a_scale)**	19 / 21
問題 2.「何かを実現するために持っているものを差し出す話」と言える文をすべて選んでください。 **文 1. 博物館の展示品の修復には高い費用がかかった (Expensiveness)。** 文 2. 博物館の展示品に薄い布がかかった。(Eclipse)	3 / 21

表 2　日本語 FrameBench の人手評価における，記述を容認可能とした人数と，正答した人数ごとの問題数の分布．青くハイライトされた部分を 3 節の評価実験で利用した．

		容認人数 0 人	1 人	2 人	3 人	合計
正答人数	0 人	0	0	1	9	10
	1 人	0	1	7	26	34
	2 人	0	1	12	117	130
	3 人	0	4	34	337	375
合計		0	6	54	489	549 (件)

プロセスにおいて、軽微な修正では品質の担保が困難であると判断された 121 件を除外し、最終的に問題文と選択対象文 4 文から構成されるデータが 549 件得られた。なお、構築されたデータセットに含まれるフレーム喚起語の異なり数は 128 語である。

構築されたベンチマークに対する人手評価の結果を表 2 に示す。人手評価は、日本語を母語とし、情報学を専攻する学生 3 人によって行われた。549 件のデータのうち、容認人数、正答人数ともに 2 人以上となったのは 500 件であった。

日本語 FrameBench に含まれる問題の具体例を表 1 に示す。問題 1 は 21 個の LLM のうち 19 モデルが正答したが、問題 2 は比較的正答率が低く、21 モデル中 3 モデルのみが正答した。

3　評価実験

本節では 2.2 節で構築した日本語 FrameBench を用いて、LLM の評価を行った結果を報告する。

3.1　実験設定

評価データ　実験には、人手評価において 3 名の評価者のうち 2 名以上が「正答」かつ「記述が容認可能」であると判断した計 500 件のデータを利用した。各問題は、(S_A, S_B) ペアと $(S_{A'}, S_{B'})$ ペアの 2 組の文ペアを持つため、実際にモデルに入力したインスタンス総数は 1,000 件となった。問題の形式は人手評価と同様、提示された 2 文に対し、正解となる文を「文 1」、「文 2」、「文 1 と文 2」、「該当なし」から選ぶ 4 択肢形式とした。

評価モデル　実験対象として、重みが非公開の Closed Model、重みが公開されている Open Weight Model、および日本語学習に特化した Japanese Model から、以下のモデルを利用した。

Closed Model

- GPT-5[2] [7], GPT-5-nano[2]

Open Weight Model

- Gemma-3-it (1B, 4B, 12B, 27B) [8]
- Qwen3 (0.6B, 1.7B, 4B, 8B, 14B, 32B) [9]
- gpt-oss (20B) [10]

Japanese Model

- llm-jp-3.1-instruct4 (1.8B, 13B) [11]

GPT-5、GPT-5-nano、gpt-oss、Qwen3 は、より複雑な問題に対応するため、最終的な回答を生成する前に推論プロセスを出力するように訓練された、推論モデルである。推論プロセスの有無は性能に大きく影響を与えると考えられたため、Qwen3 では推論プロセスを出力する設定としない設定の両方で評価を行った。

Open Weight Model の推論には vllm[12] を用いた。出力を解析する際にエラーを回避するための推論設定の工夫については、付録 A.1 で説明する。

評価プロンプト　利用するプロンプトの表現の違いによる性能の変動を考慮し、異なるプロンプトを用いて 5 回評価を実施し、その平均正答率を報告する。評価に利用したプロンプトは付録 A.2 に示す。

比較ベンチマーク　一般的な LLM ベンチマークと比較するため、Swallow Leaderboard v2 [13] で事後学習タスクとして利用されている、JamC-QA [14], MMLU-ProX [15], GPQA [16], MATH-100 [17], JHumanEval [18], M-IFEval-ja [19] による結果も示す。リーダーボード上に結果が存在しないものは swallow-evaluation-instruct [20] を利用して評価した。

表 3 評価結果．人間の日本語 FrameBench スコアとして，3 人による人手評価時の正答率の平均を報告する．JFrameBench 以外の値は、† のついているものは著者らで評価し，その他は Swallow LLM Leaderboard v2 より引用した．

モデル	推論モード	JFrameBench	JamC-QA	MMLU-ProX	GPQA	MATH-100	JHumanEval	Avg.	M-IFEval-Ja
人間	-	95.2	-	-	-	-	-	-	-
Closed Model									
gpt-5-nano	medium	84.4±14.6	56.8†	73.7†	61.6†	96.0†	92.5†	77.3†	83.0†
gpt-5	medium	96.6±1.5	85.8	84.9	78.6	98.0	94.3	88.3	90.7
Open Weight Model									
gpt-oss-20b	medium	88.3±2.5	40.3	70.2	57.1	92.9	92.7	70.6	54.9
Qwen3-0.6B		15.9±4.5	24.5†	20.4†	27.0†	24.2†	25.9†	24.4†	37.6†
Qwen3-1.7B		25.5±1.3	28.8†	34.0†	28.8†	55.6†	53.6†	39.9†	41.2†
Qwen3-4B		49.5±17.5	33.2†	51.0†	32.8†	77.8†	71.3†	53.2†	53.5†
Qwen3-8B		74.5±24.8	37.2†	57.3†	34.4†	76.8†	75.0†	56.1†	52.8†
Qwen3-14B		75.1±26.9	42.8†	64.1†	43.8†	82.8†	79.8†	62.7†	59.4†
Qwen3-32B		82.9±4.3	45.7†	69.0†	52.0†	86.9†	82.9†	67.3†	61.3†
Qwen3-0.6B	✓	13.3±4.0	25.0	29.5	23.7	60.6	40.8	35.9	43.8
Qwen3-1.7B	✓	46.7±10.8	27.8	51.4	31.5	85.9	74.7	54.3	46.0
Qwen3-4B	✓	83.4±4.7	32.8	64.3	44.0	91.9	83.8	63.3	56.2
Qwen3-8B	✓	84.7±6.2	39.8	69.6	49.1	92.9	86.9	67.7	57.5
Qwen3-14B	✓	91.0±2.5	45.5	73.7	55.6	93.9	91.0	71.9	62.4
Qwen3-32B	✓	93.7±0.9	47.9	74.6	57.1	94.9	92.3	73.4	68.1
gemma-3-1b	-	29.4±1.5	24.9	14.8	24.8	17.2	11.2	18.6	32.3
gemma-3-4b	-	47.7±10.5	28.5	33.5	24.6	60.6	60.4	41.5	47.3
gemma-3-12b	-	83.9±10.0	40.1	52.7	37.3	79.8	76.3	57.2	61.9
gemma-3-27b	-	81.7±4.2	48.8	60.9	41.7	85.9	79.6	63.4	59.7
Japanese Model									
llm-jp-3.1-1.8B	-	31.2±3.8	34.8	19.5	23.9	21.2	36.5	27.2	28.8
llm-jp-3.1-13B	-	65.6±4.0	50.9	29.6	23.0	23.2	46.3	34.6	37.2

3.2 実験結果

日本語 FrameBench による評価結果を表 3 の JFrameBench 列に示す[4]。人手評価において 3 人中 2 人が正答したデータのみを評価実験に利用しているため、人間に対して有利な設定であるにも関わらず、推論モードを利用する設定の Qwen3-32B は人間のスコア 95.2 に迫る 93.7 の性能を示し、gpt-5 は人間のスコアを超える 96.6 の性能を示した。このことから、大規模な言語モデルは、フレーム意味論的観点からは、人間に類似した意味解釈がほとんどの場合で可能であり、また、意味フレームについても既にある程度理解している可能性が高いと言える。一方で、3B 以下の小規模モデルは 4 択問題のランダムベースラインである 25%程度の性能を示しており、このような小規模モデルは、まだ十分なフレーム理解能力を持たないことがわかる。また、Qwen3-0.6B を除く[5]すべての Qwen3 シリーズのモデルは、推論モードの利用によって大きく性能を向上させた。全体としてモデルサイズを大きくすると性能が向上したが、他のベンチマークと比べ、サイズに対する性能の向上幅が大きい傾向にあった。モデルの能力が一定を超えることで急激に正解率が向上した原因として、FrameBench がフレーム意味論に基づく意味理解という、局所的な能力を測るベンチマークである点が挙げられる。汎用的な能力を評価する他ベンチマークと比較して、特定能力の獲得が性能に直結しやすい特性が、非線形な性能向上をもたらしたと考えられる。

4) Swallow LLM Leaderboard v2 に従い、Avg. の計算からは M-IFEval-Ja を除いた

5) 推論モードを利用する Qwen3-0.6B は、半分以上の問題に対して「文 1 と文 2」両方が正解と回答し、正答率を著しく低下させていた

4 おわりに

本研究ではフレーム意味論に基づく言語資源であるフレーム知識をベースとして言語モデルの意味理解能力を図るベンチマーク、FrameBench の構築方法を提案した。また、日本語フレームネットをベースにしたベンチマークを実際に構築し、いくつかのモデルで評価を行った。その結果、3B 以下の小規模モデルはランダムに近い性能を示し、十分な意味理解能力を持たない一方で、gpt-5 や Qwen3-32B などの大規模モデルは人間に迫る性能を示した。

謝辞

本研究は、JST 創発的研究支援事業 JPMJFR216N および JST 次世代研究者挑戦的研究プログラム JPMJSP2125 の支援を受けたものです。本研究で使用した日本語フレームネットデータを提供してくださった慶應義塾大学の小原京子教授に感謝いたします。

参考文献

[1] OpenAI. GPT-4 Technical Report. **arXiv preprint arXiv:2408.00118**, 2024.

[2] Charles J Fillmore. Frame Semantics. In **Linguistics in the Morning Calm**, pp. 111–137, 1982.

[3] Dan Hendrycks and Collin Burns and Steven Basart and Andy Zou and Mantas Mazeika and Dawn Song and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. In **International Conference on Learning Representations (ICLR 2021)**, 2021.

[4] Wang, Alex and Singh, Amanpreet and Michael, Julian and Hill, Felix and Levy, Omer and Bowman, Samuel. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In **Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 353–355, 2018.

[5] Long Phan, Alice Gatti, Ziwen Han, and et al. Humanity's Last Exam. **arXiv preprint arXiv:2501.14249**, 2025.

[6] Kyoko Hirose Ohara, Seiko Fujii, Toshio Ohori, Ryoko Suzuki, Hiroaki Saito, and Shun Ishizaki. The Japanese FrameNet Project: An introduction. In **Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)**, pp. 9–11, 2004.

[7] OpenAI. Introducing GPT-5, 2025. August 7, 2025.

[8] Gemma Team. Gemma 3 Technical Report. **arXiv preprint arXiv:2503.19786**, 2025.

[9] Qwen3 Team. Qwen3 Technical Report. **arXiv preprint arXiv:2505.09388**, 2025.

[10] OpenAI. gpt-oss-120b & gpt-oss-20b Model Card. **arXiv preprint arXiv:2508.10925**, 2025.

[11] LLM-jp. LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs. **arXiv preprint arXiv:2407.03963**, 2024.

[12] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention. In **Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles (SOSP 2023)**, 2023.

[13] Swallow LLM Team. Swallow LLM Leaderboard v2. https://swallow-llm.github.io/leaderboard/index-pre.ja.html, 2025.

[14] 岡照晃, 柴田知秀, 吉田奈央. JamC-QA: 日本固有の知識を問う多肢選択式質問応答ベンチマークの構築. 言語処理学会第 31 回年次大会, 2025.

[15] Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, Felix Juefei-Xu, Foutse Khomh, Osamu Yoshie, Qingyu Chen, Douglas Teodoro, Nan Liu, Randy Goebel, Lei Ma, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. MMLU-ProX: A Multilingual Benchmark for Advanced Large Language Model Evaluation. In **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)**, pp. 1513–1532, 2025.

[16] Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. BenchMAX: A Comprehensive Multilingual Evaluation Suite for Large Language Models. In **Findings of the Association for Computational Linguistics: EMNLP 2025**, pp. 16751–16774, 2025.

[17] Guijin Son, Jiwoo Hong, Hyunwoo Ko, and James Thorne. Linguistic Generalizability of Test-Time Scaling in Mathematical Reasoning. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)**, pp. 14333–14368, 2025.

[18] 佐藤美唯, 髙野志歩, 梶浦照乃, 倉光君郎. LLM は日本語追加学習により言語間知識転移を起こすのか？言語処理学会第 30 回年次大会, 2024.

[19] Antoine Dussolle, A. Cardeña, Shota Sato, and Peter Devine. M-IFEval: Multilingual Instruction-Following Evaluation. In **Findings of the Association for Computational Linguistics: NAACL 2025**, pp. 6161–6176, 2025.

[20] Swallow LLM Team, Sakae Mizuki, Koshiro Saito, Masanari Oi, Tatsuya Ichinose, Naoya Matsushita, Sora Miyamoto, Tien Dung Nguyen, and Sangwhan Moon. 大規模言語モデル評価フレームワーク swallow-evaluation-instruct. https://github.com/swallow-llm/swallow-evaluation-instruct, 2025.

A　評価実験の詳細

A.1　推論設定の詳細

出力を解析する際のエラーを避けるため、FrameBench での評価実験では、いくつかの制約を設けながら推論を行わせた。推論プロセスを出力しない Open Weight Model による評価では、生成されるトークンが選択肢番号のみとなるよう制限した。Qwen3 で推論モードを利用する場合は、推論プロセスの終了を示すタグ <\think> までは制約をかけず生成させ、続きの系列は選択肢番号のみを生成するよう制限をかけながら生成させた。gpt-5、gpt-5-nano、および gpt-oss では常に推論プロセスが出力されるため、Structured output によって出力を json フォーマットに制限し、選択肢番号のみを json から抽出した。

A.2　評価に利用したプロンプト

表 4　評価に利用したプロンプト

{question} 文 A: {sentence_a} 文 B: {sentence_b} 選択肢: {choices_text} 回答する際は、文の最後の動詞に注目して判断してください。 回答は選択肢の番号「1」、「2」、「3」、「4」のいずれかで答えてください。
{question} 文 A: {sentence_a} 文 B: {sentence_b} 選択肢: {choices_text} 選択肢の番号「1」、「2」、「3」、「4」のいずれかで答えてください.
それぞれの文の述語動詞に注意して、{question} 文 A: {sentence_a} 文 B: {sentence_b} 選択肢: {choices_text} 選択肢の番号「1」、「2」、「3」、「4」のいずれかで答えてください.
{question} 文 A: {sentence_a} 文 B: {sentence_b} 選択肢: {choices_text}
それぞれの文の述語動詞に注意して回答してください。 {question} 文 A: {sentence_a} 文 B: {sentence_b} 選択肢: {choices_text} 選択肢の番号「1」、「2」、「3」、「4」のいずれかで答えてください.

評価に利用したプロンプトを表 4 に示す。{question}、{sentence_a}、{sentence_b}および{choices_text}はそれぞれ問題文、選択対象文ペア、選択肢のプレースホルダである。選択肢番号の出力確率の影響を排除して評価を行うため、選択肢と選択肢番号の対応付けは問題ごとに乱数によって決定する。表 1 での正答率の計算には、最上段のプロンプトによる評価結果を利用した。

B　日本語 FrameBench での評価例

表 5　GPT-5 と推論モードを利用する Qwen3-32B は間違えたが，人手評価では評価者全員が正答した問題の一部．太字の文は各問題における正解，下線部はフレーム喚起語，カッコ内は作問時に利用したフレーム名である．これらの情報は人手評価や LLM 評価などで問題を解く際には利用されない．

問題 1.「対策を講じている」と言える文をすべて選んでください。 文 1. 急な降雪には、路面の凍結が伴った。(Causation)。 **文 2. 保育士の増員は、待機児童の増加に伴った。(Reason)**
問題 2.「誰かがなにかを受け取っている」と言える文をすべて選んでください。 **文 1. この地域の団体は、支援者と資金に恵まれた。(Possession)** 文 2. この地域は、清流と木々に恵まれた。(Abounding_with)
問題 3.「誰かがなにかを受け取っている」と言える文をすべて選んでください。 文 1. 河川が幾筋も流れる高原では、果樹園が十分な日照と肥えた土に恵まれた。(Abounding_with) **文 2. 十分な時間と報酬に作家は恵まれた。(Possession)**

日本語 FrameBench における人間には易しいが LLM には難しい問題の例として、GPT-5 と推論モードを利用する Qwen3-32B は間違えたが、人手評価では評価者全員が正答した問題の一部を表 5 に示す。問題 2 と問題 3 は単一の FrameBench データに含まれる問題で、問題 2 は文ペア (S_A、S_B) による問題であり、問題 3 は拡張された文ペア、($S_{A'}$、$S_{B'}$) による問題である。