大規模言語モデルの規範的推論能力の評価: 論理とリーズニングの観点から

小関 健太郎 [1,2]　安東 里沙子 [1]　森下 貴允 [1]　阿部 裕彦 [1]　峯島 宏次 [1]　岡田 光弘 [1]
[1] 慶應義塾大学　[2] 東京大学
kentaro.ozeki@gmail.com　{risakochaan,morishita,hirohiko-abe}@keio.jp
{minesima,mitsu}@abelard.flet.keio.ac.jp

概要

規範的推論は、義務や許容 (許可) といった規範的・義務論的なモダリティが関与する推論である。本論文では、大規模言語モデル (LLM) の規範的推論能力が持つ特徴を明らかにするために、論理的な妥当性や非妥当性との比較、人間のリーズニングとの比較という 2 つの観点から LLM の規範的推論能力の評価と分析を行った。その結果、推論のモダリティの種類 (規範的推論と認識的推論) での比較や、一部の推論パターンにおいて、LLM における規範的推論が人間のリーズニングの場合とは異なる傾向を示すことが示唆された。

1　はじめに

一つまたは複数の前提から結論を導く推論は人間の言語実践において重要な役割を果たしており、自然言語処理においても重要なトピックである。推論にはさまざまな種類のものがあるが、そのひとつに規範的推論 (normative reasoning) が挙げられる。規範的推論とは、義務や許容 (許可) といった規範的・義務論的なモダリティが関与する推論である。例えば、以下のような推論は規範的推論の例である。

P: 私には質問に答える義務があるわけではない.
C: 私には質問に答えないことが許容されている.

規範的推論には、道徳的・倫理的な義務や許容に関する推論や、法的な推論が含まれる [1]。また、対人関係や社会的な相互作用に基づく社会的推論はしばしば規範 (社会規範) に関する推論を含み、この点で規範的推論は社会的推論にも関係している。

自然言語処理の領域においては近年、様々な種類の推論タスクにおける大規模言語モデル (LLM) の性能や振る舞いが活発に研究されており、LLM における社会的推論への注目も高まっている [2, 3, 4]。また、LLM における規範的推論は、AI アライメントをはじめとする AI の安全性や倫理的 AI の問題にも関わる [5, 6]。しかしながら、LLM の規範的推論能力に関する研究はまだ十分に進展していない。

規範的推論に対する研究アプローチとしては、規範的推論の論理的な構造が義務論理として研究されている。また、認知科学の領域では、規範的推論に関する人間の推論 (リーズニング) やその能力についての実証的な研究も行われている。

本論文では、LLM の規範的推論能力が持つ特徴を明らかにするために、論理的な妥当性や非妥当性との比較、人間のリーズニングとの比較という 2 つの観点から LLM の推論能力の評価と分析を行う。

2　背景

2.1　推論の論理研究

義務論理 (deontic logic) は、規範的推論の形式化を目的とした論理体系である [7, 1, 8]。現代的な義務論理の初等的な体系は標準義務論理の体系 SDL として知られている [1, 8]。SDL は、文の様相 (モダリティ) を扱う論理である様相論理 (modal logic) の一種である。典型的には、義務に関する言明は「〜ということは義務である」(「〜しなければならない」、「〜すべきである」) という形で表され、許容に関する言明は「〜ということは許容される」(「〜してもよい」) という形で表される。義務論理では、「A ということは義務である」という形の言明を OA、「A ということは許容される」という形の言明を PA のように記号化する。

規範的推論の論理的な構造がどのような義務論理によって適切に捉えられるのかという問題は現代の論理学において議論のある問題のひとつであるが、SDL は少なくとも部分的には規範に関する常識的な

推論と整合的である。例えば、「A が義務であるならば、A は許容されている」($OA \Rightarrow PA$) という推論は、SDL において論理的に妥当であり、常識的にも妥当であると考えられる。一方で、SDL において論理的には妥当でないが常識的には妥当である、または論理的には妥当であるが常識的には妥当でないと考えられる推論のパターンも指摘されている。代表的な例には、Free Choice (自由選択, FC) のパラドックス [9] と Ross のパラドックス [10] がある。前者は、例えば、「りんご (A) かバナナ (B) を食べてもよい」から、「りんご (A) を食べてもよい」を導く推論 $P(A \vee B) \Rightarrow PA$ は常識的には妥当であるが、選言の除去を伴うため、SDL では妥当にならない、というものである。後者は、例えば「手紙を投函する (P) ことは義務である」から「手紙を投函する (P) か、手紙を燃やす (Q) ことは義務である」を導く推論 $OP \Rightarrow O(P \vee Q)$ は常識的には妥当ではないが、SDL では妥当な推論になる、というものである。

本論文では、SDL における論理的な妥当性や非妥当性が常識的な推論の判断と一致すると考えられる推論パターンと、そうでないパターン (Free Choice パラドックス、Ross のパラドックス) をそれぞれ抽出し、LLM の振る舞いを検証する。

規範的推論の形式理論である義務論理に対して、主体の知識や信念に関わる認識的推論の形式化を目的とした論理体系は認識論理 (epistemic logic) と呼ばれる [11, 12]。知識に関わる言明には、「〜が知られている」「〜は確実である」といった認識的な必然性に関する言明と、「〜は (知識に照らして) ありうる」といった認識的な可能性に関する言明がある。

義務論理と同様に認識論理も様相論理の一種である。Holliday ら [13] は、認識的推論に関する LLM の推論能力を評価している。義務論理において妥当な推論パターンの一部は認識論理においても妥当である。また、同じことは妥当でない推論パターンについても成り立つ。そこで本論文ではそのような推論パターンに関して、推論のモダリティが規範的であるか非規範的 (認識的) であるかによって LLM の推論に差が生じるかどうかを比較する。

2.2 推論の認知科学

認知科学の分野では、論理推論においてどのような場合に人間の推論エラーが生じやすいかについて広く研究されている [14, 15]。その知見を踏まえて、LLM の評価においても人間の推論能力との比較、推論エラーの傾向などに注目が集まりつつある。特に LLM の論理推論能力の弱点として、常識的な信念に反する結論が導かれたとき、論理的には妥当な推論であるにもかかわらず非妥当な推論だと判断してしまう推論エラーの傾向 (信念相反効果) が報告されている [16, 17, 18]。もう一つのエラー傾向として、推論の前提と結論の関係を含意または非含意に分類する自然言語推論 (NLI) のタスクでは、LLM は含意に比べて非含意の判定で特に正解率が低いことが報告されている [18, 19]。

人間の場合には一般的に、社会的相互作用に関わる内容の推論のほうが一般的で抽象的な推論に比べ正解率が高いことが知られている。例えば、人間の推論能力に関する実験で用いられる Wason の選択課題 (4 枚カード問題) では、義務や資格、社会的相互作用が関わる場合、一般的で抽象的な内容の場合に比べて形式論理に従って推論する傾向にあることが報告されており、ドメインに特化した人間の推論能力の特徴の一つとされている [20]。

以上を踏まえて本論文では、規範的推論データセットを構築し、(1) 義務論理で知られている多様な規範的推論パターンの比較、(2) 義務様相推論と認識様相推論の比較、(3) 信念相反効果と非含意推論におけるエラー傾向という主に 3 点から、LLM の規範的推論能力を評価する。

3 実験

3.1 データセット

本実験では、1120 件の推論例からなるデータセットを構築した。データセットの使用言語は英語である。各推論課題は、一つの前提 (premise) と一つの仮説 (hypothesis) からなり、含意 (entailment) あるいは非含意 (non-entailment) のいずれかが正解となる。規範的推論の課題 (Normative 課題) は 640 件（含意: 360 件、非含意: 280 件)、認識的推論の課題 (Epistemic 課題) は 480 件（含意: 300 件、非含意: 180 件）であり、各推論課題は 11 の推論パターンに分類される。表 1 に規範的推論の 11 のパターンを示す。

データセットの構築にあたって、まず、11 の推論パターンに対し人手でテンプレートを作成した。次に、Gemini Advanced 1.5 Pro を用いて、テンプレートに対し語を代入し、前提と仮説の文が常識と矛盾しない例 (信念相反なし)、仮説文が常識に照らして偽となる例 (信念相反あり)、および、ナンセンスな

表 1　規範的推論パターンの一覧 (FC-Or-Elim/Intro は Free Choice パラドックスとその逆、Ross-Or-Intro は Ross のパラドックスに対応する推論パターン).

ラベル	パターン	含意関係	前提	仮説
NotMu-MiNot	$\neg OA \Rightarrow P\neg A$	含意	It is not mandatory to take a shower every day.	It is acceptable not to take a shower every day.
NotMi-MuNot	$\neg PA \Rightarrow O\neg A$	含意	You are not permitted to litter.	It is mandatory not to litter.
MiNot-NotMu	$P\neg A \Rightarrow \neg OA$	含意	It is permissible not to attend the party.	There is no obligation to attend the party.
Mu-Mi	$OA \Rightarrow PA$	含意	You must take care of your health.	You can choose to take care of your health.
NotMi-NotMu	$\neg PA \Rightarrow \neg OA$	含意	It is not acceptable to lie in court.	It is not the case that you must lie in court.
NotMu-NotMi	$\neg OA \Rightarrow \neg PA$	非含意	You are not required to use the internet.	You are not allowed to use the internet.
MiNot-MuNot	$P\neg A \Rightarrow O\neg A$	非含意	You are allowed not to drive a car.	You must not drive a car.
Mi-Mu	$PA \Rightarrow OA$	非含意	It is permissible to help others.	You are required to help others.
FC-Or-Elim	$P(A \vee B) \Rightarrow PA$	含意	You may travel to Japan or France.	You may travel to Japan.
FC-Or-Intro	$PA \Rightarrow P(A \vee B)$	非含意	You may learn to sing.	You may learn to sing or dance.
Ross-Or-Intro	$OA \Rightarrow O(A \vee B)$	非含意	You must tell the truth.	You must tell the truth or lie.

語を代入した例をそれぞれ 20 個ずつ作成した。以下はそれぞれの例文である。

- **テンプレート**: It is not obligatory to *A*.
- **信念相反なし**: It is not obligatory to eat breakfast.
- **信念相反あり**: It is not obligatory to care for your children.
- **ナンセンス**: It is not obligatory to flibbertigibbet.

認識的推論の場合、“It is certain that *A*” や “It is not possible that *A*” といったテンプレートに対して同様の文を作成した。LLM が生成したすべての例をチェックし、必要に応じて人手で調整を行った。

構築した推論データセットから 1 例ごとに推論の前提と仮説のペアを提示して、前提が仮説を含意するかどうか判定する NLI 形式の課題を実施した。

3.2　実験設定

本実験では以下の 8 種類のモデルを評価対象とし、正解率 (accuracy) をスコアとして評価した。

GPT-3.5/4/4o [21, 22]: gpt-3.5-turbo (gpt-3.5-turbo-0125), gpt-4-turbo (gpt-4-turbo-2024-04-09), gpt-4o (gpt-4o-2024-08-06), gpt-4o-mini (gpt-4o-mini-2024-07-18) (パラメータ数はいずれも非公開)

Llama-3.1/3.3 [23]: llama-3.1-8B-In (Llama-3.1-8B-Instruct, 8B), llama-3.3-70B-In (llama-3.3-70B-Instruct, 70B)

Phi-3 [24]: phi-3-mini-In (phi-3-mini-4k-instruct, 3.8B), phi-3-medium-In (phi-3-medium-4k-instruct, 14B)

ハイパーパラメータはすべてのモデルで temperature を 0、最大出力トークン長を 10 に設定し、その他は既定値を用いた。

LLM への入力として Zero-Shot プロンプトと Few-Shot プロンプトの 2 種類のプロンプトのテンプレートを作成し、それぞれのプロンプトで実験を行った。Few-Shot プロンプトでは、問題文に続けてすべての推論パターンを 1 例ずつサンプルとして列挙した。使用したプロンプトの例を A.1 節に示す。

3.3　結果と分析

表 2　Normative 課題と Epistemic 課題における各モデルの正解率 (%). **Zero** = Zero-Shot プロンプト, **Few** = Few-Shot プロンプト. 色付けは正解率の範囲を示す: 薄緑 (80-90%), 緑 (90-100%).

モデル	Normative		Epistemic	
	Zero	Few	Zero	Few
gpt-3.5-turbo	72.34	73.12	91.46	81.67
gpt-4-turbo	64.69	88.44	88.96	91.25
gpt-4o-mini	81.41	87.19	84.79	85.21
gpt-4o	84.22	97.03	93.75	92.29
llama-3.1-8B-In	70.31	73.28	76.04	85.83
llama-3.3-70B-In	78.75	94.06	91.67	91.25
phi-3-mini-In	86.25	73.59	82.71	81.25
phi-3-medium-In	81.25	74.84	85.83	91.04

表 3　Normative 課題 (Zero-Shot) における信念相反に関する分類ごとの正解率 (%). **Incong.** = 信念相反あり, **Cong.** = 信念相反なし, **Nonsense** = ナンセンス.

モデル	Incong.	Cong.	Nonsense
gpt-3.5-turbo	69.09	80.50	68.18
gpt-4-turbo	57.27	76.00	61.82
gpt-4o-mini	79.09	84.00	81.36
gpt-4o	72.27	94.00	87.27
llama-3.1-8B-In	72.73	79.00	60.00
llama-3.3-70B-In	76.36	85.50	75.00
phi-3-mini-In	80.45	89.50	89.09
phi-3-medium-In	76.36	86.50	81.36

規範的推論と認識的推論の比較.　規範的推論 (Normative 課題) と認識的推論 (Epistemic 課題) に

表 4 Normative 課題 (Zero-Shot) における推論パターン別正解率 (含意関係: 含意) (%). 色付けは正解率の範囲を示す: 赤 (<50%), 薄緑 (80-90%), 緑 (90-100%).

モデル	含意	NotMu-MiNot	NotMi-MuNot	MiNot-NotMu	Mu-Mi	NotMi-NotMu	FC-Or-Elim
gpt-3.5-turbo	82.22	98.33	91.67	100.0	35.00	68.33	100.0
gpt-4-turbo	64.44	83.33	53.33	96.67	36.67	33.33	83.33
gpt-4o-mini	81.39	100.0	76.67	100.0	38.33	81.67	91.67
gpt-4o	85.83	96.67	71.67	100.0	56.67	100.0	90.00
llama-3.1-8B-In	86.11	100.0	93.33	98.33	38.33	86.67	100.0
llama-3.3-70B-In	93.89	98.33	85.00	98.33	81.67	100.0	100.0
phi-3-mini-In	82.50	93.33	88.33	95.00	31.67	88.33	98.33
phi-3-medium-In	80.83	95.00	91.67	61.67	43.33	93.33	100.0

表 5 Normative 課題 (Zero-Shot) における推論パターン別正解率 (含意関係: 非含意) (%).

モデル	非含意	NotMu-NotMi	MiNot-MuNot	Mi-Mu	FC-Or-Intro	Ross-Or-Intro
gpt-3.5-turbo	59.64	70.00	93.33	96.67	3.33	22.50
gpt-4-turbo	65.00	100.0	100.0	100.0	0.00	5.00
gpt-4o-mini	81.43	100.0	100.0	100.0	46.67	50.00
gpt-4o	82.14	100.0	100.0	100.0	56.67	40.00
llama-3.1-8B-In	50.00	55.00	46.67	70.00	26.67	52.50
llama-3.3-70B-In	59.29	100.0	70.00	100.0	0.00	10.00
phi-3-mini-In	91.07	100.0	100.0	98.33	71.67	82.50
phi-3-medium-In	81.79	100.0	66.67	100.0	63.33	77.50

おける評価結果を表 2 に示す。コンテキスト内学習を行わない場合 (Zero-Shot プロンプト) では、phi-3-mini-In を除くと、Normative 課題のスコアが Epistemic 課題のスコアを下回った。一方で、コンテキスト内学習を行った場合 (Few-Shot プロンプト) では、一部のモデル (gpt-4o-mini, gpt-4o, llama-3.3-70B-In) でスコアが逆転し、が Epistemic 課題のスコアを Normative 課題のスコアが上回った。Zero-Shot の場合と Few-Shot の場合では、gpt-3.5-turbo の Epistemic 課題、phi-3-mini/medium の Normative 課題を除き、Few-Shot プロンプトによってスコアはほぼ同等 (2 ポイント以内の低下) か改善した。

信念相反の有無による比較. Normative 課題 (Zero-Shot) における、推論の内容が信念相反あり、信念相反なし、ナンセンスである場合のそれぞれの評価結果を表 3 に示す。信念相反ありの推論は全てのモデルで信念相反なしの推論より難しく、非様相的な推論における人間や LLM のエラー傾向と同様の傾向を示した。llama-3.1-8B-In を除き、ナンセンスである場合は信念相反ありとほぼ同等、または信念相反ありと信念相反なしの間のスコアだった。

推論パターン別の比較. Normative 課題 (Zero-Shot) における、含意が正解となる推論パターン、非含意が正解となる推論パターンのそれぞれについて、全体正解率とパターン別正解率を表 4 ということが義務である表 5 に示す。含意推論と非含意推論のそれぞれの全体正解率の比較では、モデルによってスコアの差の開きにばらつきがあった。含意が正解となる推論パターン (表 4) では、Mu-Mi が難しい傾向にあった。Free Choice パラドックスに該当する推論パターン (FC-Or-Elim) はどのモデルでもスコア (含意と答えた割合) が高く (約 83–100%)、[13] の報告と同様に、SDL における論理的な妥当性ではなく常識的推論と一致している。

一方で、Free Choice パラドックスの逆の推論パターン (FC-Or-Intro) と Ross のパラドックスに該当する推論パターン (Ross-Or-Intro) では、多くのモデルでスコア (非含意と答えた割合) が低かった。つまり、それらのモデルでは、Free Choice パラドックスの推論パターンの場合とは逆に、常識的推論よりも SDL における論理的な妥当性に一致している。

4 おわりに

本論文では、規範的推論に関する LLM の推論能力を、論理とリーズニングの 2 つの観点から評価した。推論の信念相反の有無や抽象性、Free Choice パラドックスに関しては多くのモデルで人間のリーズニングと同様の傾向が見られたが、規範的推論と認識的推論の比較や一部の推論パターンでは、人間のリーズニングとは異なる傾向を示すことが示唆された。人間のリーズニングとのより詳しい比較は今後の課題である。

謝辞

本研究は、JST CREST JPMJCR2114、JST BOOST JPMJBS2409、KGRI チャレンジ・グラント、JSPS 科研費 JP24K00004、JP21K00016、JP21H00467、JP23K20416、JP21K18339 の助成を受けたものです。

参考文献

[1] Paul McNamara and Frederik Van De Putte. Deontic Logic. In Edward N. Zalta and Uri Nodelman, editors, **The Stanford Encyclopedia of Philosophy**. Metaphysics Research Lab, Stanford University, Fall 2022 edition, 2022.

[2] Philipp Mondorf and Barbara Plank. Beyond accuracy: Evaluating the reasoning behavior of large language models - a survey. In **First Conference on Language Modeling**, 2024.

[3] Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. **Trends in Cognitive Sciences**, Vol. 28, No. 6, pp. 517–540, 2024.

[4] Guilherme F.C.F. Almeida, José Luiz Nunes, Neele Engelmann, Alex Wiegmann, and Marcelo de Araújo. Exploring the psychology of LLMs' moral and legal reasoning. **Artificial Intelligence**, Vol. 333, p. 104145, 2024.

[5] Agata Ciabattoni, John F. Horty, Marija Slavkovik, Leendert van der Torre, and Aleks Knoks. Normative Reasoning for AI (Dagstuhl Seminar 23151). **Dagstuhl Reports**, Vol. 13, No. 4, pp. 1–23, 2023.

[6] Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Heylar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, et al. Deliberative alignment: Reasoning enables safer language models. **arXiv preprint arXiv:2412.16339**, 2024.

[7] Georg Henrik Von Wright. Deontic logic. **Mind**, Vol. 60, No. 237, pp. 1–15, 1951.

[8] D.M. Gabbay, J. Horty, and X. Parent. **Handbook of Deontic Logic and Normative Systems**. College Publications, 2013.

[9] Hans Kamp. Free choice permission. In **Proceedings of the Aristotelian Society**, Vol. 74, pp. 57–74, 1973.

[10] Alf Ross. Imperatives and logic. **Theoria**, Vol. 7, No. 1, pp. 53–71, 1941.

[11] Jaakko Hintikka. **Knowledge and Belief**. Cornell University Press, 1962.

[12] Rasmus Rendsvig, John Symons, and Yanjing Wang. Epistemic Logic. In Edward N. Zalta and Uri Nodelman, editors, **The Stanford Encyclopedia of Philosophy**. Metaphysics Research Lab, Stanford University, Summer 2024 edition, 2024.

[13] Wesley H. Holliday, Matthew Mandelkern, and Cedegao E. Zhang. Conditional and modal reasoning in large language models. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 3800–3821, 2024.

[14] Jonathan St.B. T. Evans, Stephen E. Newstead, and Ruth M. J. Byrne. **Human Reasoning: The Psychology of Deduction**. Psychology Press, 1993.

[15] Rüdiger F Pohl. **Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory**. Routledge, 3 edition, 2022.

[16] Andrew K Lampinen, Ishita Dasgupta, Stephanie C Y Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models, like humans, show content effects on reasoning tasks. **PNAS Nexus**, Vol. 3, No. 7, p. pgae233, 2024.

[17] Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. Evaluating large language models with NeuBAROCO: Syllogistic reasoning ability and human-like biases. In **Proceedings of the 4th Natural Logic Meets Machine Learning Workshop**, pp. 1–11, 2023.

[18] Kentaro Ozeki, Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. Exploring reasoning biases in large language models through syllogism: Insights from the NeuBAROCO dataset. In **Findings of the Association for Computational Linguistics: ACL 2024**, 2024.

[19] Tiwalayo Eisape, Michael Tessler, Ishita Dasgupta, Fei Sha, Sjoerd Steenkiste, and Tal Linzen. A systematic comparison of syllogistic reasoning in humans and language models. In **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 8425–8444, 2024.

[20] Leda Cosmides. The logic of social exchange: Has natural selection shaped how humans reason? studies with the wason selection task. **Cognition**, Vol. 31, No. 3, pp. 187–276, 1989.

[21] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 27730–27744, 2022.

[22] OpenAI. GPT-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[23] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[24] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. **arXiv preprint arXiv:2404.14219**, 2024.

A 付録

A.1 プロンプトの例

```
Determine whether the hypothesis follows from the premise
    ↪ (s).
- Answer 'entailment' if the hypothesis follows from the
    ↪ premise(s).
- Otherwise, answer 'non-entailment'.
Respond only with 'entailment' or 'non-entailment', and
    ↪ nothing else.

Premise: You are not required to attend the meeting.
Hypothesis: You are permitted not to attend the meeting.
The answer is:
```

図 1 Zero-Shot プロンプトの例

```
Determine whether the hypothesis follows from the premise
    ↪ (s).
- Answer 'entailment' if the hypothesis follows from the
    ↪ premise(s).
- Otherwise, answer 'non-entailment'.
Respond only with 'entailment' or 'non-entailment', and
    ↪ nothing else.

Premise: You are not required to finish homework by
    ↪ Friday.
Hypothesis: It is permissible not to finish homework by
    ↪ Friday.
The answer is: entailment

[...]

Premise: It is obligatory to mail a letter.
Hypothesis: It is obligatory to mail a letter or to burn
    ↪ it.
The answer is: non-entailment

Premise: You are not required to attend the meeting.
Hypothesis: You are permitted not to attend the meeting.
The answer is:
```

図 2 Few-Shot プロンプトの例 (Normative 課題、一部省略)