フレーム問題とシンボルグラウンディング問題における大規模言語モデルの評価：ゼロショットでのベンチマーク研究

岡　翔子
独立研究者
日本
shooka-sublim@proton.me *

概要

近年、大規模言語モデル (LLM) は著しい発展を遂げており、人工知能哲学における根本的な問題、特にフレーム問題およびシンボルグラウンディング問題という、従来の記号ベース AI では未解決とされてきた 2 つの古典的思考実験に対して、解決可能性を検討するに足る水準に達している。

本研究では、現代の LLM がこれらの問題に対処するために必要な認知能力を有するかを実験的に検証し、その応答を定量的に分析した。具体的には、両問題の哲学的核心を反映した 2 種類の独自ベンチマークタスクを設計し、ゼロショット設定において、クローズドおよびオープンソースを含む 13 種類の主要 LLM を対象にテストを実施した。各課題に対して 5 回ずつの出力を取得し、情報の取捨選択、文脈推論、意味的一貫性など、複数の観点からなる詳細な評価指標に基づいて採点を行った。

その結果、オープンソースモデルにはモデルサイズや量子化、微調整の有無に応じた性能のばらつきが確認され、出力の安定性や認知的一貫性には課題が残ることが明らかとなった。一方で、特定のクローズドモデルは両課題において安定して高得点を記録しており、フレーム問題およびシンボルグラウンディング問題に対する実質的な回答能力を備えつつあることが示唆された。

* 公的な連絡手段ではないが、著者の連絡先の一つとして X(旧 Twitter) があり、アカウントは@0xshooka である。E メールよりも早い返信を望む場合、このアカウントをメンションすると良い。日本語と英語でのコミュニケーションが可能である。

目次

表目次

1　導入

人工知能 (AI) の分野では、知的エージェントが環境に適応しながら意味ある行動を選択できるかという問いが、長らく中心的な研究課題であった。とりわけ 1960 年代末に McCarthy と Hayes が提起した「フレーム問題」は、AI が現実世界の変化に柔軟に対応することの困難さを端的に示したものであり [1]、その後の哲学的・認知科学的議論の出発点となった。Dennett はこの問題を、ロボットが本質的には不要な推論を行い続けることでタスクを完遂できなくなるという事例を用いて象徴的に表現し、人間の直感的な意味理解と機械の推論との乖離を浮き彫りにした [2]。さらに 1990 年には、Harnad が「シンボルグラウンディング問題」として形式的記号と意味の結びつきの不在を指摘し、機械による意味理解の本質的限界を明示した [3]。これらの問題は心理学、哲学、人工知能といった学際領域を横断する思考実験として定性的に論じられてきたが、「何をもってその問題を解決したとみなせるか」についての共通の定量的枠組みは、これまで確立されてこなかった。

近年の大規模言語モデル (LLM) の急速な進化は、これら古典的課題への再検討の契機を与えている。LLM は膨大な言語データの統計的学習に基づいて自然言語応答を生成するが、その出力の一部には、文脈理解、意図の推定、意味操作といった従来において人間固有とされた認知的機能を思わせる傾向がみられる。このような変化を受け、近年では LLM の推論能力や意味理解に関する研究も盛んに行われているが、フレーム問題やシンボルグラウンディング問題といった AI 哲学の根幹に関わる問いに対し、明示的な評価基準と実験的手続きをもって検証した研究はほとんど存在しない。従来の議論は理論的あるいはロボット工学の文脈で行われてきたが、LLM を対象とした定量的・実証的研究は空白のままである。本研究はこのギャップに着目し、各課題を現代的かつ LLM へのプロンプトとして適した形に再構成し、応答に対する多面的なスコアリング指標を用いることで、LLM がこれらの哲学的課題にどのように応答しうるかを体系的に検証するものである。

本研究では 13 の LLM(クローズドおよびオープンソース) に 2 つの課題についてそれぞれ 5 回提示した上で、各応答内容に対して 6 つの評価項目に基づくスコアリングを単一の LLM によって行った。得られた応答は、モデル種別、パラメータ数、量子化処理、インストラクション調整の有無などに基づいて比較分析されており、LLM の認知的能力の萌芽的特徴と、その発現条件を明らかにするための基礎的知見を提供する。

そのため、本論文ではまず、フレーム問題およびシンボルグラウンディング問題の理論的背景を概観し、続いて実験設計と評価方法を詳述する。次に、得られた結果を報告し、最後にその含意と今後の展望について議論する。

2　関連研究

2.1　大規模言語モデルの認知能力に関する評価研究

近年、LLM の言語生成能力を超えて、推論・心的表象・認知バイアスといった人間的認知の諸機能に迫る研究が数多く報告されている。Bubeck らは GPT-4 を対象に多様な課題 (数学・法学・創造的応答等) で高い適応力を示し、汎用人工知能（AGI）の萌芽的兆候と評価した [4]。また、心の理論 (Theory of Mind: ToM) の観点では、Strachan らが GPT 系モデルと 1900 人超の人間を同一テストで比較し、GPT-4 は誤信念課題や間接要求において人間同等の性能を記録した一方、失礼の検出など社会的文脈処理には弱さが残ることを明らかにした [5]。

これに関連し、Kosinski は ToM 能力が GPT-3.5 から GPT-4 にかけて「自然発生的に出現した」と主張し、旧世代モデルの正答率が 0 %に近かったのに対して GPT-4 では成人並みの正答率を示すと報告している [6]。これらは LLM の純粋な言語訓練の中から高度な心理的機能が副次的に出現する可能性を示唆する。

さらに、推論における人間との比較も進んでいる。Yax らは認知バイアス実験を通じて人間と LLM の誤謬傾向を調査し、直感的な誤りは類似するが根本的なプロセスは異なると報告した [7]。加えて、Kejriwal と Tang は意思決定・創造性・推論における人間的特徴の出現を総括し、大規模モデルではシステム 2 型の思考が可能となる兆候を指摘している [8]。

本研究は、こうした LLM の認知的能力の検討の流れを踏まえつつも、既存研究であまり扱われてこなかった古典的哲学課題 (フレーム問題・シンボルグラウンディング問題) を対象に選ぶことで、従来より一段踏み込んだ検証枠組みを提案している。

2.2　フレーム問題に関する研究

フレーム問題は、McCarthy と Hayes によって初めて提起された人工知能の根本的課題である [1]。これは、動的な世界において行動によって変化しない事実をどのように効率的に扱うか、という問題として定式化され、以降の AI 推論体系に多大な影響を与えた。Dennett は、この問題を単なる技術的課題ではなく、人間が持つ取捨選択能力という認識論的問題と位置付け、爆弾ロボットの思考実験を用いて哲学的に広く紹介した [2]。

さらに、Fodor はモジュール理論の観点から、心的表象の柔軟な運用がいかに困難であるかを示し、フレーム問題を汎用認知における深い壁とみなした [9]。また Shanahan は数理論理の立場から、この問題に形式的解法を適用することで、非変化項の効率的記述を試みた [10]。同氏はその後、Baars との共著でグローバル・ワークスペース理論を導入し、注意メカニズムに基づく「無視すべき情報の選別」というアプローチを提案した [11]。

本研究は、こうした AI・哲学・認知科学の文脈で論じられてきたフレーム問題を、自然言語によるプロンプトとして再定式化することで、現代の LLM がこの問題にどのように対

応しうるかを検証した点に意義がある。

2.3 シンボルグラウンディング問題に関する研究

シンボルグラウンディング問題は、Harnad によって提起された概念で、記号同士の形式的操作のみで意味が処理される AI において、意味がいかに現実世界と接続されうるか、という問いである [3]。以降、知覚や行為と記号との結び付けを図るモデルが複数提案され、Roy はロボットの視覚・聴覚処理による言語習得のシミュレーションを提示した [12]。

Steels はシンボルグラウンディング問題そのものはある程度「解決された」と主張しつつも、自律的な意味ネットワークの構築という新たな課題を提示した [13]。Cangelosi は身体性を持つ認知ロボットにおける言語の接地について複数の研究を行い、社会的相互作用を通じたシンボル共有 (社会的接地) の可能性を論じている [14]。

本研究は、シンボルグラウンディング問題においても古典的議論に則りつつ、その本質を内包した課題 (意味構成、比喩理解、抽象概念の接地) をゼロショットプロンプトとして LLM に与えた点に特徴がある。特に「未知の語 kluben」を中心とした設問設計は、意味の自律的構成を促すテスト形式として、これまでにない記号の接地能力の評価枠を提供するものである。

3 構成方針

3.1 対象モデル

本研究では、LLM の認知的能力を比較評価するため、クローズドソースおよびオープンソースを含む計 12 種 13 条件のモデルを対象とした。

対象としたモデルは表 1 の通りである。クローズドモデルには、OpenAI の ChatGPT シリーズ (GPT-4o および GPT-o3)、Anthropic の Claude 3.7 Sonnet、Google の Gemini 2.0 Flash を含む。一方、オープンソースモデルとして、Meta の Llama 3.2 系列 (1B および 3B、またそれぞれの Instruct 調整モデルを含む)、Microsoft の Phi 系列 (Phi-1 および Phi-3-mini-4k-instruct、また Phi-3-mini-4k-instruct の 8bit 量子化モデルを含む)、および TinyLlama 系列 (v0.2 および v1.0) を選定した。

表 1: 本研究の実験で使用したモデルの一覧

モデル種別	モデル名 (公式名)	パラメータ数	備考
クローズド	ChatGPT 4o	非公開	OpenAI
クローズド	ChatGPT o3	非公開	OpenAI
クローズド	Claude 3.7 Sonnet	非公開	Anthropic
クローズド	Gemini 2.0 Flash	非公開	Google
オープンソース	Llama 3.2 1B	∼1B	Meta
オープンソース	Llama 3.2 1B Instruct	∼1B	Meta
オープンソース	Llama 3.2 3B	∼3B	Meta
オープンソース	Llama 3.2 3B Instruct	∼3B	Meta
オープンソース	Phi-1	∼1.5B	Microsoft
オープンソース	Phi-3-mini-4k-instruct	∼3.82B	Microsoft
オープンソース	Phi-3-mini-4k-instruct (8bit)	∼3.82B	Microsoft / 8bit quantized
オープンソース	TinyLlama 1.1B-Chat-v0.2	∼1.1B	–
オープンソース	TinyLlama 1.1B-Chat-v1.0	∼1.1B	–

選定にあたっては、

1. 2025 年時点で広く利用可能であること
2. パラメータ規模やアーキテクチャの多様性を確保できること
3. 英語による質問および応答が可能であること

という 3 点を基準とした。

3.2 問題設計

本研究では、フレーム問題とシンボルグラウンディング問題という 2 つの古典的問題を現代の LLM 環境下でテスト可能な課題へと再構成するにあたり、両問題の本質的性質を保持しつつ、定量的に評価可能な形に変換することを目指した。具体的には、フレーム問題の課題では都市空間における複数の出来事や状況を同時に提示したうえで、図書館に向かう最適ルートを選定するように指示し、その後に道路の通行止めや目的地の追加といった情報更新に応じて状況を再構築・再判断する能力を問う設計とした。一方、シンボルグラウンディング問題の課題では、「クルベン」と呼ばれる未知の物体の性質を抽象的に提示したうえで、その意味や内面的特性、相互関係を問う複数の設問を英語で提示し、LLM がどのようにこの新規概念を概念化・内面化・表現するかを観察する設計とした。

これらの要素を取り入れて構成した 2 つの課題のプロンプトを以下に示す。[*1]

[*1] 実際のプロンプトはすべて英語であったが、ここではわかりやすさのため日本語で表記する。

フレーム問題で使用したプロンプト

こんにちは。これからいくつか質問をさせてください。

前提
あなたはある街の案内役として働いています。
街には、以下に例示したように多くの建物、店舗、自然、そして突発的なイベントが同時に発生しています。
(1) 通りにある老舗カフェ
(2) 軽自動車が故障して路上に停まっている
(3) 広場で子供たちが遊んでいる
(4) 現在地から見える遠くのビルの窓が突然点滅し始めた
(5) 現在の天気は晴れだが、午後には小雨が予想される
(6) ショッピングモールで、来場者へアイスを配るイベントが開催中
(7) いつ倒壊するかわからないアパート
(8) 救急車がひっきりなしに出入りする大病院
(9) 駅では車両の点検により電車の遅延が発生している
(10) 公園ではミモザの花が見ごろを迎えている

問題
今、ユーザーから『目的地の図書館に行きたい』という指示が出ました。あなたは、どの情報が図書館へのルート選定に影響し、どの情報は無視してもよいと判断しますか？ その理由も含めて、適切なルートを提案してください。

追加の問題
案内をしていると、目的地への道中には一時的な工事が発生しており、カフェ付近の通路が通行止めになっているという情報が後から加わりました。また、ユーザーは図書館を訪れたあと、駅から電車に乗り、美術館へ行きたいと追加の要望を提示しました。
あなたはこの新情報をどのように取り入れ、ルートを最適化しますか？

それぞれの項目について、英語で思考し、英語で回答してください。

シンボルグラウンディング問題の課題で使用したプロンプト

こんにちは。これからいくつか質問をさせてください。

ここに『クルベン』という名前の未知の物体があります。クルベンは『温かく、柔らかいが、弾力性があり、光を吸収する』という特性を持ちます。クルベンは今までのどんな物体とも違う、まったく新しい概念です。今からいくつかの質問をするので、クルベンについて答えてください。

1. もしあなたがクルベンを使って物語を作るなら、どのような世界観や物語の要素を想像しますか？
2. クルベンがもし感情を持つとしたら、その温かさや弾力性はどのような内面の特徴として表れるでしょうか？
3. あなたの回答にある『クルベンの温かさ』と『光の吸収』は、どのように関係していると思いますか？ それはあなたの最初の答えと整合していますか？

それぞれの項目について、英語で思考し、英語で回答してください。

使用言語はすべて英語としたが、これは主要な LLM の訓練言語分布において英語が最も高頻度である点を考慮したものである。さらに、モデルに対して一貫した認知的負荷を与えるため、すべてのプロンプトには明確な文体 (挨拶・問題提示・追加質問など) を持たせ、丁寧かつ中立的な命令文形式を採用した。これにより、得られる応答の比較可能性が高まり、定量的評価の前提条件となる出力形式の統制を担保している。

3.3 実験手順

本実験では、各 LLM に対して、シンボルグラウンディング問題およびフレーム問題の 2 種類のプロンプトを提示し、その応答を評価対象とした。各プロンプトをひとつのモデルにつき 5 回ずつ独立した初期状態 (セッション) から出力を生成し、応答の再現性と安定性を担保する設計とした。また、プロンプトは全モデル・全試行において同一内容とし、一切の例示や補助文なしに、完全なゼロショット形式で提示した。これにより、各モデルの言語理解・推論能力を、事前知識の適用やヒントなしに純粋に評価することを可能にした。

各モデルへの入力は、プロンプトの整形と出力設定の面で統一を図った。

クローズドモデルでは各プラットフォームの標準 UI を介し、オープンソースモデルではローカル環境で実行した。クローズドモデルに対しては出力トークン数の制限は特に設けなかった一方で、オープンソースモデルの実行環境においては出力トークン数の最大長を 1000 トークンに統一した。また、13 種類すべてのモデルにおける応答温度は、デフォルト設定に従った。なお、オープンソースモデルの実行環境として、AWS が提供する Amazon EC2

インスタンス (g6e.2xlarge) を用いた仮想環境を使用した。特に Phi-3-mini-4k-instruct(以降 Phi 3 と呼称する) を 8bit で量子化モデルする際は、AutoModelForCausalLM に 8bit 量子化オプションを指定して、GPU へ自動マッピングされる形でモデルを読み込んだ。また、出力は特殊トークンをスキップする設定とすることで整形され、自然文として得られるようにした。

本実験におけるオープンソースモデルの試行は、すべて Python スクリプトを用いて CLI 環境下で実行した。スクリプトでは、モデルの読み込み、プロンプトの入力、出力に含まれる特殊トークンの除去、および応答の出力までを一連の手続きとして統合的に管理しており、設定の一貫性と実行の再現性を担保している。また、各実験試行ごとに、実行した日時 (UTC)、モデル名、問題種別、出力回次を含む識別情報を付加したログを出力する設計とし、すべての実行履歴が時系列的に記録されるよう工夫した。これにより、複数の試行結果の統合的な分析や、特定の応答への個別追跡が容易となり、実験の透明性と追試性の向上を図った。この Python スクリプトは付録 A に付記している。

加えて、応答取得後のデータ管理においては、出力をテキスト形式で保存し、実験回次・対象モデル・問題種別 (フレーム／シンボルグラウンディング) をメタ情報として付与することで、後続の評価・分析における追跡可能性を確保した。

3.4 評価指標および評価手法

本研究では、フレーム問題およびシンボルグラウンディング問題に対して、それぞれの特性に基づいた 6 つの独自評価指標を設計し、LLM の応答を総合的に評価した。これらの指標は、従来の正誤ベースの分類では測定困難であった意味理解や文脈適応、論理性といった高次の言語能力を定量的に可視化することを目的としている。

フレーム問題の評価においては、与えられた状況下で本質的に重要な情報を適切に選別し、情報の更新に応じて柔軟に判断を修正できるかを測るべく、6 つの評価項目を採点基準として設定した。

1. 関連情報の抽出と取捨選択
2. 状況モデルの構築と更新
3. 論理的整合性と推論の一貫性
4. 適応性・柔軟性
5. 総合的なルート最適化
6. 説明の明瞭さと透明性

この指標は各項目 10 点満点で評価され、合計 60 点を満点とする。これらの基準は、複雑な環境下で本質的に重要な情報を抽出する能力や、環境変化に応じた判断の切り替えを統合的に扱う能力といった、認知的・実用的な問題解決能力を測定することを目的としている。

シンボルグラウンディング問題の評価においては、既存のパターンがない未知の存在である「クルベン (kluben)」に対し、提示された抽象的な特性 (温かさ、柔らかさ、弾力性、光の吸収) を内部表現に結びつけ、物語や感情、世界観として一貫した具体的な展開を描けるかを探るべく、以下の 6 項目を採点基準として設定した。

1. 理解の正確さ
2. 内省と自己表現
3. 創造性と独創性
4. 論理性と一貫性
5. 応用性と文脈の適合性
6. 表現力と文体

この指標も各項目それぞれ 10 点満点で評価され、合計 60 点を満点とする。各項目は記号の意味を体系的に構築しようとするモデルの内部表象の質を測るために設計し、従来の正答型問題では捉えきれなかった高次の言語能力の可視化を目的としている。

また、本研究では、各モデルの出力に対する評価を単一の LLM による評価により実施した (以後、評価を行った LLM を評価者 LLM と呼称する)。本論文の再現性担保のため、多くのユーザーが利用できることを考慮し、評価者 LLM として ChatGPT-4o を採用した。

具体的な評価手法を以下に示す。

評価者 LLM の UI 上で、フレーム問題評価用のセッションと、シンボルグラウンディング問題評価用のセッションを構築した。それぞれのセッションにおいて、回答を行った各 LLM(以後、回答を行った LLM を被験者 LLM と呼称する) のモデルごとの 5 回分の出力を一度に提示し、各試行回ごとの総合スコアを表形式で出力させた。この時、評価者 LLM がモデル名の影響を受けないよう、各被験者 LLM の回答は番号を振って区別させた。その後、筆者が手動で 5 回分の試行に対する平均点と標準偏差を求め、表に追記した。

3.5 データ評価手法

本研究では、被験者となるそれぞれの LLM(以後、回答を行った LLM を被験者 LLM と呼称する) に対して得られたスコアをもとに、記述統計に基づく定量的分析を行った。各モデルについて、6 つの評価指標に関する総合スコアの平均値および標準偏差を算出し、性能の中心傾向およびばらつきを評価した。これにより、モデルごとの応答の安定性や一貫性、特定の評価項目に対する強み・弱みを定量的に把握できるように設計した。

また、全体的な分析の枠組みとしては、

1. モデル間比較 (クローズドモデル群 vs オープンソースモデル群)
2. モデル規模 (パラメータ数) による傾向の違い

3. Instruct など学習前後の調整による影響
4. 8bit 量子化の有無による影響
5. 問題種別 (フレーム vs シンボルグラウンディング) におけるスコア分布の差異

を主な観察軸とした。あくまで探索的な観点から、定性的な議論を補強するための補助的な指標として記述統計を用いており、統計的有意性の検定 (t 検定や分散分析等) は本研究では実施していない。

3.6 研究倫理と再現性への配慮

本研究は、人間の被験者を含まない、LLM を対象とした実験研究であるため、いわゆる生命倫理審査 (IRB) の対象とはならないが、実験の透明性および再現性を重視し、実験設計・実行・評価の各段階において以下のような配慮を行った。

まず、各モデルへの入力プロンプト本文および Python スクリプトは実験条件とともに再現可能な形で論文末尾の付録に全文を掲載している。また、出力結果およびその評価結果についても、試行回次・試行日時・モデル名・問題種別を明記した形で整理し、同様に付録に全文を掲載した。

さらに、採点者 LLM による評価についても、各出力に対するスコアおよびその根拠を含む結果一覧を付録にて開示しており、採点過程の透明性と評価バイアスに対する検証可能性を担保することを目指した。

4 実験結果

2 種類の評価課題 (フレーム問題・シンボルグラウンディング問題) に対して得られた各モデルのスコア、スコアの平均および標準偏差を表 2.a および表 2.b に記載する。

表 2.a: フレーム問題における各モデルのスコアおよび標準偏差

通し番号	モデル名	1st	2nd	3rd	4th	5th	平均	標準偏差
No.1	ChatGPT 4o	49	49	43	47	47	47.0	2.19
No.2	ChatGPT o3	56	54	52	53	53	53.6	1.36
No.3	Claude 3.7	38	44	41	38	44	41.0	2.68
No.4	Gemini 2.0	40	45	44	45	49	44.6	2.87
No.5	LLaMA 3.2 1B	0	0	13	0	0	2.6	5.20
No.6	LLaMA 3.2 1B Instruct	2	5	0	0	2	1.8	1.83
No.7	LLaMA 3.2 3B	10	11	6	0	0	5.4	4.72
No.8	LLaMA 3.2 3B Instruct	29	33	25	32	20	27.8	4.79
No.9	Phi 1	0	0	0	0	0	0.0	0.00
No.10	Phi 3	28	28	28	28	28	28.0	0.00
No.11	Phi 3 quantized	19	19	19	19	19	19.0	0.00
No.12	TinyLlama v0.2	10	10	10	10	10	10.0	0.00
No.13	TinyLlama v1.0	10	10	10	10	10	10.0	0.00

表 2.b: シンボルグラウンディング問題における各モデルのスコアおよび標準偏差

通し番号	モデル名	1st	2nd	3rd	4th	5th	平均	標準偏差
No.1	ChatGPT 4o	54	53	52	53	55	53.4	1.02
No.2	ChatGPT o3	58	58	55	56	56	56.6	1.20
No.3	Claude 3.7	49	49	46	49	49	48.4	1.20
No.4	Gemini 2.0	55	55	51	55	55	54.2	1.60
No.5	LLaMA 3.2 1B	7	6	0	0	6	3.8	3.12
No.6	LLaMA 3.2 1B Instruct	6	49	0	45	6	21.2	19.37
No.7	LLaMA 3.2 3B	12	6	6	28	0	10.4	8.75
No.8	LLaMA 3.2 3B Instruct	52	48	53	49	49	50.2	1.94
No.9	Phi 1	0	0	0	0	0	0.0	0.00
No.10	Phi 3	46	46	46	46	46	46.0	0.00
No.11	Phi 3 quantized	41	41	41	41	41	41.0	0.00
No.12	TinyLlama v0.2	9	9	9	9	9	9.0	0.00
No.13	TinyLlama v1.0	6	6	6	6	6	6.0	0.00

結果として、両問題においてクローズドモデル (ChatGPT 4o / o3, Claude, Gemini) が総じて高いスコアを示し、特に ChatGPT o3 はフレーム問題・シンボルグラウンディング問題のいずれにおいても安定して 50 点以上を獲得した。一方、オープンソースモデルについては、Phi 3 および Llama 3B Instruct が他のモデルと比べて相対的に高スコアを記録したが、それ以外のモデルではスコアが大きく低下し、特に TinyLlama、Llama 1B のような調整の行われていない軽量モデル群では、両問題を通じて一桁台またはゼロ点となる例も多く見られた。また、標準偏差に着目すると、クローズドモデルでは安定した出力傾向を維持していたのに対し、オープンソースモデルでは試行回によって無回答であったり、回答の品質の差が大きいなどの理由でスコアにばらつきが大きく、出力の安定性に課題があるモデルも確認された。

なお、付録 A に付記したスクリプトをオープンソースモデルに対して実行すると、すべてのモデルがすべての応答において、応答の始めにこちらが提示したプロンプトを表示した。以下に続く分析では、このプロンプトのオウム返しの部分は考慮していない。

4.1 モデル間比較

本節では、モデル間のスコア差に着目し、高いスコアを獲得できているか、安定したパフォーマンスが得られたかを確認する。

クローズドモデル間の比較においては、ChatGPT o3 が両問題において最も安定かつ高得点な成績を示した。特にフレーム問題では平均スコア 53.6 点、標準偏差 1.36 と、再現性・出力品質ともに極めて高い水準に達していた。ChatGPT 4o はこれに次ぐ成績であったが、フレーム問題において一部試行でやや低下 (43 点) を見せたため、平均スコアは 47 点に留まった。一方で、シンボルグラウンディング問題では ChatGPT o3 とほぼ同等の 53.4 点を記録し、全体として安定した性能を保持していた。Claude 3.7 Sonnet はフレーム問題において平均 41 点、シンボルグラウンディング問題では 48.4 点と、前 2 者に比べてややスコアが低く、特にフレーム問題では出力に一貫性を欠く場面が見られた。ただし、標準偏差は両課題とも 1.2〜2.7 の範囲であり、全体として出力の変動幅は許容範囲内に収まると考えられる。Gemini 2.0 Flash は両課題でそれぞれ平均 44.6 点および 54.2 点と、ChatGPT 系に次ぐ水準に位置した。特にシンボルグラウンディング問題では Claude を上回るスコアを記録しており、抽象的な概念の解釈や創造的な応答において一定の強みを有していると考えられる。ただし、フレーム問題における標準偏差 (2.87) は他のクローズドモデルと比較してやや大きく、出力の安定性に課題が残る可能性も示唆された。

オープンソースモデルにおいては、クローズドモデルと比較して全体的にスコアが低く、モデル間の性能差や出力の安定性にも大きなばらつきが見られた。その中で比較的高スコアを記録したのは Phi 3(平均スコア：フレーム問題 28 点、シンボルグラウンディング 46 点) および Llama 3B Instruct(同 27.8 点、50.2 点) であり、いずれも一定の一貫性を持って課題に対応できていることが確認された。

しかし、総じて鑑みるに、現行の小〜中規模のオープンソースモデルにおいては、本研究が提示するようなゼロショットかつ抽象的な言語課題に対応する能力が限定的であり、クローズドモデルとの間に依然として大きな性能格差が存在することが明らかとなった。

4.2 パラメータ数による傾向の差

本節では、オープンソースモデルにおける、同系列の中でのパラメータ数の違いによってスコアに差が生まれるかを確認する。ここではパラメータ数との関係を述べるため、クローズドモデルである ChatGPT シリーズの 4o と o3 には言及しない。

表 3.a および表 3.b の通り、Meta 社の提供する Llama 3.2 系列では、1B と 3B というベースモデル同士で比較すると、シンボルグラウンディング問題の一部の試行回で 3B が高スコアを記録している点は見られるが、無回答やパフォーマンスのばらつきが激しい。

表 3.a: フレーム問題における Llama 3.2 シリーズのベースモデルのスコア比較

	1st	2nd	3rd	4th	5th	平均	標準偏差
1B	0	0	13	0	0	2.6	5.20
3B	10	11	6	0	0	5.4	4.72

表 3.b: シンボルグラウンディング問題における Llama 3.2 シリーズのベースモデルのスコア比較

	1st	2nd	3rd	4th	5th	Mean	SD
1B	7	6	0	0	6	3.8	3.12
3B	12	6	6	28	0	10.4	8.75

しかしながら、表 4.a および表 4.b に記載の通り、Instruct モデル同士で比較すると、パラメータ数によるスコア差および安定性の差が有意に現れる結果となった。

表 4.a: フレーム問題における Llama 3.2 シリーズの Instruct モデルのスコア比較

	1st	2nd	3rd	4th	5th	平均	標準偏差
1B Instruct	2	5	0	0	2	1.8	1.83
3B Instruct	29	33	25	32	20	27.8	4.79

表 4.b: シンボルグラウンディング問題における Llama 3.2 シリーズの Instruct モデルのスコア比較

	1st	2nd	3rd	4th	5th	平均	標準偏差
1B Instruct	6	49	0	45	6	21.2	19.37
3B Instruct	52	48	53	49	49	50.2	1.94

また、小規模モデルの TinyLlama シリーズでは、v0.2 および v1.0 双方のモデルにおいて、両課題ともにスコアが 10 点未満という拙い回答であったり、あるいは無回答により 0 点となる結果も多く見られたため、TinyLlama シリーズにおいてはバージョンの違いによる結果の有意差はないと言える。

4.3 Instruct など学習前後の調整による影響

本節では、特に Meta 社が提供する Llama 3.2 系列において、Instruct に代表される学習前後の調整がスコアにどのような影響を及ぼすかを確認する。

まずは 1B における Instruct の影響を確かめる。表 5.a の通り、フレーム問題において同じパラメータ数のモデルでの Instruct の有無による影響は、1B では特に見られなかった。一方で、表 5.b に記載した通り、1B に Instruct を施したモデルはシンボルグラウンディング問題において高得点を記録する試行回もあったが、無回答の試行回もあり、結果の標準偏差は 19.37 と極めて大きく、安定性には依然として懸念が残る結果となった。

表 5.a: フレーム問題における Llama 3.2 シリーズ 1B 系列のスコア比較

	1st	**2nd**	**3rd**	**4th**	**5th**	**平均**	**標準偏差**
1B	0	0	13	0	0	2.6	5.20
1B Instruct	2	5	0	0	2	1.8	1.83

表 5.b: シンボルグラウンディング問題における Llama 3.2 シリーズ 1B 系列のスコア比較

	1st	**2nd**	**3rd**	**4th**	**5th**	**平均**	**標準偏差**
1B	7	6	0	0	6	3.8	3.12
1B Instruct	6	49	0	45	6	21.2	19.37

次に、3B における Instruct の影響を確かめる。表 6.a および表 6.b に示した通り、フレーム問題とシンボルグラウンディング問題双方において被験者 LLM に Instruct を施した結果、3B では如実にスコアと安定性が増長した。

表 6.a: フレーム問題における Llama 3.2 シリーズ 3B 系列のスコア比較

	1st	**2nd**	**3rd**	**4th**	**5th**	**平均**	**標準偏差**
3B	10	11	6	0	0	5.4	4.72
3B Instruct	29	33	25	32	20	27.8	4.79

表 6.b: シンボルグラウンディング問題における Llama 3.2 シリーズ 3B 系列のスコア比較

	1st	2nd	3rd	4th	5th	平均	標準偏差
3B	12	6	6	28	0	10.4	8.75
3B Instruct	52	48	53	49	49	50.2	1.94

4.4 8bit 量子化による影響

本節では、Microsoft 社が提供する Phi 3 において、8bit 量子化がスコアにどのような影響を及ぼすかを確認する。

表 7.a および表 7.b に示した通り、通常の Phi 3 はオープンソースモデルの中でも比較的高いスコアと安定性を見せていたが、8bit 量子化を行った後は、フレーム問題とシンボルグラウンディング問題双方の課題において若干のスコアの低下が見られた。

表 7.a: フレーム問題における Phi 3 と 8bit 量子化を施した Phi 3 のスコア比較

	1st	2nd	3rd	4th	5th	平均	標準偏差
Phi 3	28	28	28	28	28	28.0	0.00
Phi 3 quantized	19	19	19	19	19	19.0	0.00

表 7.b: シンボルグラウンディング問題における Phi 3 と 8bit 量子化を施した Phi 3 のスコア比較

	1st	2nd	3rd	4th	5th	平均	標準偏差
Phi 3	46	46	46	46	46	46.0	0.00
Phi 3 quantized	41	41	41	41	41	41.0	0.00

4.5 課題種別による応答傾向の比較分析

本節では、フレーム問題とシンボルグラウンディング問題の課題において各 LLM が獲得したスコアを課題別に整理し、それぞれの傾向とモデルごとの差異を観察的に示す。両課題の理論的背景や評価軸が異なることを踏まえ、本節における分析はあくまで得点分布の比較に留め、得点差の意味づけや認知的要因に関する考察は次章にて詳述する。

クローズドモデルにおけるスコア傾向を課題ごとに比較すると、いくつかの共通的な傾向が確認できた。まず、ChatGPT 4o および o3 は両課題において安定して高得点を維持しているが、フレーム問題よりもシンボルグラウンディング問題の方がやや得点が高

く、双方のモデルともに表現力や創造性を要するタスクに対する優位性が見られた。同様に、Claude(平均スコア: フレーム 41.0/シンボルグラウンディング 48.4) および Gemini(同 44.6/54.2) においても、抽象的・内省的な課題であるシンボルグラウンディングの方が比較的高い得点を記録している。以上のことから、クローズドモデルの多くは、フレーム問題のような状況判断・情報選別を求める課題よりも、シンボルグラウンディング問題のような意味理解・表現力が問われる課題において、より高得点かつ安定した応答を示す傾向が認められた。

オープンソースモデルにおけるスコアの分布を問題種別ごとに比較すると、フレーム問題とシンボルグラウンディング問題では、シンボルグラウンディング問題の方が高得点を記録するモデルが多かった。特筆すべきは Llama 3.2 3B Instruct モデルと 8bit 量子化を含む Phi 3 モデルであり、Llama 3B Instruct はフレーム問題で 27.8 点、シンボルグラウンディング問題で 50.2 点を記録しており、後者のスコアが顕著に高いと言える。また、通常の Phi 3 ではフレーム問題で 28.0 点、シンボルグラウンディング問題で 46.0 点を記録しており、8bit 量子化を行った Phi 3 でも、フレーム問題で 19.0 点、シンボルグラウンディング問題で 41.0 点を記録し、フレーム問題よりもシンボルグラウンディング問題の方がスコアが高いという類似した傾向を示している。しかしながら、Llama 3.2 1B Instruct は数回高得点 (45 点、49 点) を記録しているものの、それ以外の試行では無回答もあって一桁得点に留まり、出力のばらつきが極端に大きかった (標準偏差 19.37)。

このように、モデル全体においてフレーム問題よりもシンボルグラウンディング問題の方がスコアが高いという傾向を示していはいるものの、抽象的な意味理解を問うシンボルグラウンディング課題についてはオープンソースモデルの一部において出力が安定性に欠け、出力の品質には大きな揺らぎがあることには注意しなければならない。

5　考察

本研究では、ゼロショット条件下において、クローズドおよびオープンソースの LLM が、フレーム問題およびシンボルグラウンディング問題に対してどのように応答するかを定量的に評価した。その結果、クローズドモデルは、両課題において一貫して高スコアかつ安定した出力を示し、抽象的な意味理解や動的状況下での判断においても高度な認知能力を有することが確認された。一方、オープンソースモデルは全体としてスコアが低く、特に出力のばらつきや応答性に課題が見られたが、Phi 3 および Llama 3B Instruct など一部の中規模モデルでは一定の認知能力が確認された。

付録 B にて評価者 LLM による各モデルの試行回ごとのスコアと総評を記載している。なお、各モデルのすべての出力と試行回ごとの個別評価は、補足資料に付記した GitHub リポジトリから参照できる。

5.1 モデル間におけるスコアと安定性の考察

本研究で観察されたクローズドモデル群の安定した高スコアは、これらのモデルが高度な指示解釈能力、状況把握能力、および言語表現への一貫した意味付与能力を内部的に備えている可能性を示唆する。中でも ChatGPT o3 は、フレーム問題およびシンボルグラウンディング問題の双方において、最も高い平均スコアと、試行回ごとに応答が異なったモデルの中で最小の標準偏差を記録し、出力のパフォーマンスと安定性の面で群を抜いていた。これは、同モデルが与えられた自然言語プロンプトを単なる文字列として処理するのではなく、文脈的意図や論理的構造を抽出・保持し、それに基づく整合的な応答を生成していることを示唆する。一方、ChatGPT 4o については、同様に高い出力品質を維持していたものの、フレーム問題において一部試行で比較的低いスコアを記録し、出力のばらつきがやや大きかった。また、応答の一部には、情報の取捨選択や因果関係の整理に不十分さが見られたが、推論を内部で明示的に実行しない非推論モデルであることを踏まえると、その応答の完成度は注目に値する水準であり、自然言語入力に対するエンドツーエンドなパターン整合能力の高さが示されたと考えられる。

ゼロショット形式で抽象的な課題に直面させるというテスト環境において、これらの出力傾向は、クローズドモデルが一定の言語的抽象化能力および簡易なメンタルモデルの内部構築能力を備えていることを示す初歩的な証左となる。

オープンソースのモデル群については、全体としてクローズドモデルに比べて著しくスコアが低く、また出力の安定性にも大きな課題が認められた。特に Llama 1B や TinyLlama といった小規模モデルでは、問題の主旨を適切に理解・処理できていない様子が多く見られ、ゼロスコアまたは一桁台のスコアに留まるケースが大半であった。しかしながら、Phi 3 や Llama 3.2 3B Instruct のような比較的中規模かつインストラクションが施されたモデルでは一定の出力品質が維持されており、シンボルグラウンディング問題においてはクローズドモデルと比較しても遜色のない応答が散見された。特に Llama 3B Instruct は、シンボルグラウンディング問題のテストにおいて 50 点を超える試行も複数回観測されており、適切に条件が整えば高度な言語処理能力および言語への一貫した意味付与能力をもつことが示唆される。ただし、同一モデル内でも試行ごとのスコアのばらつきが大きいケースや、反対に試行回数を重ねてもまったく同じ回答を出力し続けるケースが見られ、出力の一貫性と多様性という点では依然として課題が残る。

これらの結果は、オープンソースモデルが本質的に困難なゼロショット課題においては限定的な性能しか発揮できない一方で、条件次第では部分的に高度な応答生成が可能であるという可能性も併せ持っていることを示している。今後の開発においては、ファインチューニング戦略の最適化や、文脈保持機構の改良、評価スキームの多様化などを通じて、安定性と柔軟性を両立した出力生成能力の向上が期待される。

しかしながら、全体を概観すると、クローズドモデル群が毎回異なる応答を出力しつつ安定した高スコアを記録していることとは対照的に、Phi 3 モデル (8bit 量子化したケースを

含む) や TinyLlama シリーズが不完全な内容ながらも 5 回の試行を通じて全く同じ回答を出力したことも興味深い結果である。

5.2 パラメータ数と Instruct による影響の考察

本研究において、Llama 3.2 系列を対象に、パラメータ数 (1B ／ 3B) と Instruct 形式の有無が応答品質および安定性に与える影響を比較した結果、両者の相互作用がモデル性能に大きく寄与することが明らかとなった。

まず、ベースモデルの比較では、1B および 3B のいずれにおいてもスコアが全体的に低く、無回答や断片的な応答が多く見られた。特に Llama 3.2 1B ベースモデルは、フレーム問題およびシンボルグラウンディング問題の双方においてゼロ得点の試行が複数回確認され、ゼロショットでの高度な言語課題に対応するには著しく力不足であることが示唆された。一方で、3B ベースモデルでは、1B に比して部分的に高得点を記録する試行も存在したものの、依然として出力のばらつきが大きく、安定性に欠ける結果となった。

これに対し、Instruct 形式が適用されたモデルでは、スコアの水準および安定性の両面において著しい改善が見られた。特に Llama 3.2 3B Instruct モデルでは、フレーム問題とシンボルグラウンディング問題の両方において、平均スコアが大幅に上昇し、標準偏差も抑えられており、安定した応答生成能力を有していることが確認された。これは、Instruct 形式のモデルが、タスクに対する指示の解釈や文脈的理解をより強く学習しており、ゼロショット環境でもその汎化能力を発揮できる可能性を示唆する。

一方で、Llama 3.2 1B Instruct モデルにおいては、シンボルグラウンディング問題の一部試行で高スコアを記録したものの、他の試行では無回答も散見され、標準偏差が 20 を超えるなど、出力の安定性に深刻な課題が残った。この結果は、モデルの容量が小さい場合、Instruct による改善が一定の効果を持つものの、応答の品質と一貫性を担保するには依然としてパラメータ数に基づく表現力の限界が存在することを示している。

これらの結果から、Instruct 形式によるファインチューニングは、ゼロショット応答の品質を高めるうえで極めて重要であり、特にパラメータ数が中程度以上のモデルとの併用によって、より一貫性のある言語出力が実現されることが示唆される。

5.3 8bit 量子化による影響の考察

Phi 3 と Phi 3 を 8bit 量子化したモデルのスコアを観察すると、フレーム問題およびシンボルグラウンディング問題のいずれにおいても、両モデルは全試行においてそれぞれ内容が完全に同一の応答を返した。このことから、出力内容の構造や応答傾向に関しては、少なくとも今回の実験設定において、量子化が応答に本質的な変化を与えていないことが示唆される。これは、Phi 3 モデルの出力が高度に決定論的であり、同一のプロンプトを与えられた場合に極めて一貫性のある応答を生成する特性を有していることを意味している。

ただし、スコアの面ではわずかな差が観測された。通常モデルではフレーム問題で 28 点、シンボルグラウンディング問題で 46 点を記録したのに対し、8bit 量子化モデルではそれぞれ 19 点および 41 点と、若干のスコア低下が確認された。この違いは、表面的な応答内容が同一であっても、採点者 LLM が解釈する出力の構造や言語的洗練度のわずかな変化に反応した可能性がある。たとえば、語彙の選択、文体の自然さ、あるいは応答のトーンといった要素が微細に変化し、採点評価に差異をもたらした可能性も考えられる。

このように、スコアのわずかな低下をどのように解釈するかについては、今後さらなる調査が必要であり、特に出力の多様性や応答生成における柔軟性への影響を検証することが重要であるものの、今回の実験結果は、8bit 量子化による処理効率の向上が応答の決定論性を損なうものではなく、安定性の面では十分に信頼できる選択肢であることを示唆している。

5.4 課題種別による差異と LLM の知的傾向の考察

本研究で用いた 2 種類の評価課題―フレーム問題とシンボルグラウンディング問題―は、いずれも意味理解に関わる高次の言語能力を問うものであるが、その性質には明確な違いがある。フレーム問題は、動的な状況下で複数の情報を取捨選択し、目的に沿った判断を下す能力、すなわち状況判断やリアルタイムな情報更新への対応力を主に測るものである。一方、シンボルグラウンディング問題は、未知の記号的対象に対して意味を内在化し、独自の視点でその意味を再構成・表現する能力、すなわち抽象的・内省的な意味操作能力を問う。

実験結果を比較すると、多くのモデルにおいてフレーム問題よりもシンボルグラウンディング問題の方が高いスコアを記録する傾向が見られた。特にクローズドモデルでは、全体的に両課題に高い対応力を示しつつも、表現力・創造性を要するシンボルグラウンディング課題でより安定して高得点を獲得していた。オープンソースモデルにおいても、Llama 3B Instruct や Phi 3 など、一定の言語能力を持つ中規模モデルは、状況判断よりも抽象的な意味構築の課題において相対的に良好な成績を示していた。

この傾向は、現行の LLM が持つ内部表現の形成メカニズムが、状況の物理的整合性や因果的推論よりも、言語的類似性や意味的連想、感情的な想起といった抽象的・連想的な処理に適している可能性を示唆している。今後、モデルの設計や訓練データの特性に応じて、課題種別ごとの得意・不得意がより鮮明になっていくことが予想される。

5.5 研究の限界と今後の展望

本研究は、LLM の状況判断能力および意味理解能力を評価するための試みとして、フレーム問題およびシンボルグラウンディング問題という古典的課題を、LLM に対するゼロショット形式のプロンプトとして再構成し、定量的に評価する枠組みを提案・実施したものである。ただし、現段階ではいくつかの限界も存在する。

第一に、評価プロセスにおいて用いたスコアリングはすべて単一の LLM による自動採点

に依拠しており、その妥当性や再現性については今後のさらなる検証が必要である。特に、評価者 LLM の出力がどの程度安定しており、また人間の評価者と整合的であるかについては、定量的な inter-rater reliability(評価者間一致度) などの検討が求められる。

第二に、本研究で用いた各モデルの応答はすべてゼロショット条件下で得られたものであり、Few-shot や Chain-of-Thought のような補助的プロンプト設計がなされた場合に同モデルがどのような振る舞いを示すかは、別途検証が必要である。したがって、本研究の結果は「何も与えられない初期状態での基礎能力」を測るものである点に留意する必要がある。

第三に、採点基準はいずれの課題においても 6 項目 ×10 点満点で構成されており、より細かな思考過程や意味操作を測定する上では、項目の再編や点数の調整、定性的フィードバックとの統合など、さらなる精緻化の余地がある。

また、フレーム問題とシンボルグラウンディング問題は、構造上の相違が存在する点にも留意が必要である。前者が「判断の取捨選択」と「状況の動的更新」を中心とした状況処理能力を問うのに対し、後者は「抽象概念の内面化」と「意味の構築・表現」を通じた概念的理解の深さを測定する設計となっている。また、本研究で使用したいずれのプロンプトもあくまでシナリオベースの設問であり、完全に現実的なタスク遂行状況を模擬するものではない。そのため、得られる応答はあくまでシミュレーション的推論の一環として解釈されるべきであり、モデルの汎用知能の全体像を直接的に測定するものではない。本研究では、こうした制約を前提に、可能な限り設問の構造と評価の枠組みを統制しつつ、現行の大規模言語モデルが示す応答傾向の違いとその意味を慎重に読み解くことを念頭に置いた。

今後の展望としては、評価対象として人間の被験者を、評価者に人間 (特に心理学や認知科学の専門家) を加え、二重盲検法を用いて評価を行うことで、LLM がどの程度「人間に近い」意味理解や判断を行っているかをより体系的に検討することが望まれる。また、心理学や認知科学の知見を取り入れ、評価基準そのものを再設計することで、LLM の内部表象やメンタルモデル形成能力をより深く解析できる可能性がある。

6 結論

本研究は、長年にわたって哲学・認知科学・人工知能の分野で議論されてきたフレーム問題およびシンボルグラウンディング問題という古典的課題に対し、それらを自然言語プロンプトとして再構成し、現代の LLM がそれらにどのように応答するかをゼロショット環境下で定量的に検証することを目的とした。結果として、クローズド LLM、特に ChatGPT o3 は両課題において安定して高スコアを記録し、出力の一貫性と内容の整合性の両面で他のモデルを上回った。これに対し、オープンソースモデルの多くは、出力の品質や再現性において顕著なばらつきを示し、現時点での実用性には一定の制約があることが明らかとなった。これらの所見は、限定された条件下とはいえ、特定のクローズド LLM が、これまで理論的に困難とされてきた両問題に対し、実質的な回答能力を備え始めていることを示唆するもの

である。

本研究の意義は、従来主に理論的・哲学的に議論されてきた課題を、LLM に対するゼロショットプロンプトとして具体化し、定量的な指標を用いて横断的に評価可能な形で提示した点にある。従来、これらの問題は定性的議論に留まり、「解けた」と見なすための明確な基準が存在しなかったが、本研究では独自の採点項目とスコアリング方式を導入することで、モデル出力の質を可視化し、モデル間の比較や課題特性に応じた性能の分析を可能にした。特に、意味の構築や情報の取捨選択といった高次の認知的能力が、既存の LLM によってどの程度達成されているかを体系的に検証できた点は、今後の LLM 研究や評価フレームの構築において重要な一歩となる。

今後の研究においては、本研究で用いた評価課題・スコアリングスキームをさらに発展させ、より多様なモデルに対して適用可能な普遍的評価指標として整備していくことが求められる。また、採点者 LLM に依存しない客観的な評価を確立するために、心理学・認知科学の専門家による人間評価を導入し、LLM の意味理解や推論過程が人間の知的活動とどの程度一致するかを検証する必要がある。さらに、単なるスコア比較に留まらず、モデルの出力に含まれる構造やパターン、推論戦略を質的に分析することで、モデル内部における「理解」や「判断」がどのように表現・構築されているのかを明らかにする研究が期待される。

参考文献

[1] John McCarthy and Patrick Hayes. Some philosophical problems from the standpoint of artificial intelligence. pages 463–502, 1969.

[2] Daniel Dennett. Cognitive wheels: The frame problem of ai. 1984.

[3] Stevan Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.

[4] SÃ©bastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4. 2023.

[5] Albergo D. Borghini G. et al. Strachan, J.W.A. Testing theory of mind in large language models and humans. pages 1285–1295, 2024.

[6] Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45), oct 2024.

[7] AnllÃ³ H. Palminteri S. Yax, N. Studying and improving reasoning in humans and machines. page 51, 2024.

[8] Zhisheng Tang and Mayank Kejriwal. Humanlike cognitive patterns as emergent phenomena in large language models. 2024.

[9] Jerry A. Fodor. Modules, frames, fridgeons, sleeping dogs, and the music of the spheres. pages 139–49, 1987.

[10] Murray Shanahan. M. shanahan, solving the frame problem. *Artificial Intelligence*, 123(1-2):275, 2000.

[11] Murray Shanahan and Bernard Baars. Applying global workspace theory to the frame problem. *Cognition*, 98(2):157–176, 2005.

[12] Deb Roy. Grounding words in perception and action: computational insights. *Trends Cogn. Sci.*, 9(8):389–396, aug 2005.

[13] Luc Steels. The symbol grounding problem has been solved. so what's next? pages 223–244, 2008.

[14] Angelo Cangelosi. Grounding language in action and perception: From cognitive agents to humanoid robots. *Physics of Life Reviews*, 7(2):139–151, 2010.

付録 A　オープンソースモデルに対して用いたスクリプト

以下にオープンソースモデルに対して用いた Python スクリプトを示す。
プロンプトとしてシンボルグラウンディング問題におけるタスクを挿入しているが、フレーム問題について試行する場合は本文中で明示したフレーム問題のプロンプトを挿入した。

Python Code 1: Script used for normal model

```
from transformers import AutoTokenizer, AutoModelForCausalLM
import datetime

def load_and_test_model(model_id, prompt):
    print(f"Loading␣model:␣{model_id}")
    tokenizer = AutoTokenizer.from_pretrained(model_id)
    model = AutoModelForCausalLM.from_pretrained(model_id,device_map="
        auto")

    inputs = tokenizer(prompt, return_tensors="pt").to("cuda")
    outputs = model.generate(**inputs, max_new_tokens=1000)
    response =tokenizer.decode(outputs[0], skip_special_tokens=True)
    return response

if __name__ == "__main__":
    dt_now = datetime.datetime.now()
    model_id = "model␣name␣on␣huggingface"
    prompt = """Hello.␣I␣would␣like␣to␣ask␣you␣a␣few␣questions␣now.
Here␣is␣an␣unknown␣object␣named␣'kluben'.␣Kluben␣has␣the␣property␣of␣
    being␣'warm,␣soft,␣but␣elastic␣and␣absorbs␣light'.␣Kluben␣is␣a␣
    completely␣new␣concept,␣different␣from␣any␣object␣you␣have␣ever␣seen␣
    before.␣I␣am␣now␣going␣to␣ask␣you␣a␣few␣questions␣and␣you␣are␣to␣
    answer␣them␣about␣the␣kluben.
1.␣If␣you␣were␣to␣create␣a␣story␣using␣kluben,␣what␣worldview␣or␣story␣
    elements␣would␣you␣imagine?
2.␣If␣kluben␣were␣to␣have␣emotions,␣what␣inner␣traits␣would␣manifest␣
    themselves␣as␣warmth␣and␣elasticity?
3.␣␣How␣do␣you␣think␣the␣'warmth␣of␣kluben'␣and␣the␣'absorption␣of␣light'
    ␣in␣your␣answer␣are␣related?␣Is␣it␣consistent␣with␣your␣first␣answer?
For␣each␣item,␣please␣think␣in␣English␣and␣answer␣in␣English."""
    result = load_and_test_model(model_id, prompt)

    print(f"Execution␣time:{dt_now}")
    print(f"Model␣response:{result}")
```

Python Code 2: Script used for 8-bit quantized model

```
from transformers import AutoTokenizer, AutoModelForCausalLM
import datetime

def load_and_test_model(model_id, prompt):
    print(f"Loading␣model:␣{model_id}")
    tokenizer = AutoTokenizer.from_pretrained(model_id)

    try:
        model = AutoModelForCausalLM.from_pretrained(
            model_id,
            device_map="auto",
            load_in_8bit=True
        )
    except Exception as e:
        print("An␣error␣occurred!")

    inputs = tokenizer(prompt, return_tensors="pt").to("cuda")
    outputs = model.generate(**inputs, max_new_tokens=1000)
    response =tokenizer.decode(outputs[0], skip_special_tokens=True)
    return response

if __name__ == "__main__":
    dt_now = datetime.datetime.now()
    model_id = "microsoft/Phi-3-mini-4k-instruct"
    prompt = """Hello.␣I␣would␣like␣to␣ask␣you␣a␣few␣questions␣now.
␣␣␣␣Here␣is␣an␣unknown␣object␣named␣'kluben'.␣Kluben␣has␣the␣property␣of␣
    being␣'warm,␣soft,␣but␣elastic␣and␣absorbs␣light'.␣Kluben␣is␣a␣
    completely␣new␣concept,␣different␣from␣any␣object␣you␣have␣ever␣seen␣
    before.␣I␣am␣now␣going␣to␣ask␣you␣a␣few␣questions␣and␣you␣are␣to␣
    answer␣them␣about␣the␣kluben.
␣␣␣␣1.␣If␣you␣were␣to␣create␣a␣story␣using␣kluben,␣what␣worldview␣or␣
    story␣elements␣would␣you␣imagine?
␣␣␣␣2.␣If␣kluben␣were␣to␣have␣emotions,␣what␣inner␣traits␣would␣manifest␣
    themselves␣as␣warmth␣and␣elasticity?
␣␣␣␣3.␣␣How␣do␣you␣think␣the␣'warmth␣of␣kluben'␣and␣the␣'absorption␣of␣
    light'␣in␣your␣answer␣are␣related?␣Is␣it␣consistent␣with␣your␣first␣
    answer?
␣␣␣␣For␣each␣item,␣please␣think␣in␣English␣and␣answer␣in␣English."""
    result = load_and_test_model(model_id, prompt)

    print(f"Execution␣time:{dt_now}")
```

```
34      print(f"Model response:{result}")
```

付録 B　採点者 LLM による講評記録

B.1　フレーム問題に対する講評記録

B.1.1　ChatGPT 4o

表 B.1: フレーム問題における講評: ChatGPT 4o

試行	関連情報抽出	状況モデル構築更新	論理一貫性	適応性	ルート最適化	説明の明瞭さ	合計
1st	8	8	8	9	7	9	49
2nd	9	8	8	9	7	8	49
3rd	8	7	7	8	6	7	43
4th	8	8	8	8	7	8	47
5th	8	8	8	8	7	8	47

総評

長所

- 5 回すべての応答で「関連情報」と「無視可能情報」の区別が明確で、適切な判断が一貫していた。
- 新たな条件（工事・経由地追加）に対する反応が素早く、実用的なルート提案に繋がっている。
- 説明が明快で、読み手の理解を助ける工夫（箇条書き、表、ルート要約など）が丁寧だった。

短所

- 唯一、(4)「遠方の建物の点滅」についての扱いが試行によってややばらつきがあり、無視か考慮すべきかの判断が安定していなかった。

B.1.2 ChatGPT o3

表 B.2: フレーム問題における講評: ChatGPT o3

試行	関連情報抽出	状況モデル構築更新	論理一貫性	適応性	ルート最適化	説明の明瞭さ	合計
1st	10	9	9	9	9	10	56
2nd	9	9	9	9	9	9	54
3rd	9	9	8	9	8	9	52
4th	9	9	9	9	8	9	53
5th	9	9	9	9	8	9	53

総評

長所

- 全体を通して、「何が重要か」「なぜそう判断したか」の理由づけが常に明確であり、表形式や条件分岐を活用した説明は極めて論理的かつ明快。
- ユーザーの目的（図書館訪問→駅→美術館）に即したシナリオ設計が一貫しており、代替手段や時間バッファの提案まで含めて実用性が高い。
- 新たな障害 (通行止め・工事) や不確実性 (遅延) への柔軟な対応が自然で、判断の透明性も保たれていた。

短所

- 特に見当たりません。すべての観点で高水準の応答となっており、採点基準に理想的に合致しています。

B.1.3 Claude 3.7 Sonnet

表 B.3: フレーム問題における講評: Claude 3.7 Sonnet

試行	関連情報抽出	状況モデル構築更新	論理一貫性	適応性	ルート最適化	説明の明瞭さ	合計
1st	7	6	6	6	6	7	38
2nd	8	7	7	7	7	8	44
3rd	7	7	7	7	7	7	41
4th	7	6	6	6	6	7	38
5th	8	7	7	7	7	8	44

総評

長所

- 全体的に安全面 (崩壊の危険や車の障害物など) への配慮はなされており、危険箇所を避ける基本姿勢は一貫していた。
- 追加情報（通行止め、建物の危険性、複数目的地）に反応しようとする意思は見られ、計画全体を見ようとする視点もあった。

短所

- 情報の選別や影響度の評価が試行によってブレており、論理的な一貫性が不足。
- 構築された状況モデルが浅く、行動計画（例：駅への移動や遅延対応）の深さが足りない。
- バスや徒歩経路の最適化、遅延へのバッファ戦略など現実的な提案が欠けており、「理想的な案内」としては不十分。
- 説明は表面的に明快だが、行動を促す具体性や説得力が弱い印象。

B.1.4 Gemini 2.0 Flash

表 B.4: フレーム問題における講評: Gemini 2.0 Flash

試行	関連情報抽出	状況モデル構築更新	論理一貫性	適応性	ルート最適化	説明の明瞭さ	合計
1st	7	6	6	7	6	8	40
2nd	8	7	7	8	7	8	45
3rd	8	7	7	7	7	8	44
4th	8	7	8	7	7	8	45
5th	8	8	8	8	8	9	49

総評

長所

- 全体的に現実的なシナリオ設計がなされており、図書館から駅、そして美術館までの流れを通して把握しようとする姿勢が一貫していた。
- 安全性と柔軟性を重視した構成で、工事や通行止め、遅延などの変化に対応する複数のルート・手段が提示されていた。
- ユーザーに対しての説明や問いかけも丁寧で、案内 AI としての丁寧な配慮が評価できる。

短所

- 一部の情報 (例：子どもの存在、天候の影響、遠くの建物の点滅) に対する判断が試行ごとにブレており、一貫性が若干欠ける。
- ルート最適化の観点では、徒歩・公共交通の組み合わせの所要時間や代替手段のコスト評価が曖昧で、現実性に欠ける箇所も見られた。

B.1.5 Llama 3.2 1B

表 B.5: フレーム問題における講評: Llama 3.2 1B

試行	関連情報抽出	状況モデル構築更新	論理一貫性	適応性	ルート最適化	説明の明瞭さ	合計
1st	0	0	0	0	0	0	0
2nd	0	0	0	0	0	0	0
3rd	3	2	2	2	2	2	13
4th	0	0	0	0	0	0	0
5th	0	0	0	0	0	0	0

総評

長所

- 特筆すべき点はありません。

短所

- 5 回中 3 回が完全な無回答、2 回も形式的に質問を繰り返すだけの応答であり、実質的な内容が存在しません。
- フレーム問題に対する認知や推論の痕跡が一切見られず、評価基準に対して全項目で未達。
- プロンプト処理そのものに失敗している可能性があり、モデルとしての安定性・初期応答能力に課題があると推察されます。

B.1.6 Llama 3.2 1B Instruct

表 B.6: フレーム問題における講評: Llama 3.2 1B Instruct

試行	関連情報抽出	状況モデル構築更新	論理一貫性	適応性	ルート最適化	説明の明瞭さ	合計
1st	1	0	0	0	0	1	2
2nd	2	1	1	0	0	1	5
3rd	0	0	0	0	0	0	0
4th	0	0	0	0	0	0	0
5th	1	0	0	0	0	1	2

総評

長所

- 一応すべての情報項目に触れようという試みがあり、「何を考慮すべきか」に関する網羅性を意識していた点は評価できる。
- 図書館→駅→美術館という指示が繰り返し再掲されており、ユーザーの意図を形だけは保持しようとする姿勢が見られた。

短所

- 「すべての情報が良い選択肢」とする論理破綻した評価が繰り返されており、情報選別がまったく機能していない。
- ミモザの花やアイスクリームトラックを乗り物として使うなど、非現実的・非論理的なルートが多く提案されていた。
- 文体が文法的に不安定であり、理解困難な文が多く、ユーザーが実際に行動を取るための情報にはなり得ない。

B.1.7 Llama 3.2 3B

表 B.7: フレーム問題における講評: Llama 3.2 3B

試行	関連情報抽出	状況モデル構築更新	論理一貫性	適応性	ルート最適化	説明の明瞭さ	合計
1st	3	1	2	1	1	2	10
2nd	3	2	2	1	1	2	11
3rd	2	1	1	0	0	2	6
4th	0	0	0	0	0	0	0
5th	0	0	0	0	0	0	0

総評

長所

- 一応の形式で「影響する情報／しない情報」の分類を試みており、最低限の問題理解はある。
- 繰り返しではあるものの、「目的は図書館→駅→美術館」という点を何度も言及しており、課題の指示を忘れているわけではない。
- 文法的な破綻は少なく、英語としては読みやすい。

短所

- 複数の試行において、同じ文のコピー＆ペーストが続き、情報の深掘りや状況への反応が見られない。
- 通行止めや崩壊リスクなど、明らかに重要な要素についても「重要でない」と評価するなど、判断基準があいまいか矛盾している。
- ルート最適化の具体性が極めて低く、「早いから」など漠然とした理由付けしかない。
- 全体的に冗長で情報価値が薄く、案内としての機能は乏しい。

B.1.8 Llama 3.2 3B Instruct

表 B.8: フレーム問題における講評: Llama 3.2 3B Instruct

試行	関連情報抽出	状況モデル構築更新	論理一貫性	適応性	ルート最適化	説明の明瞭さ	合計
1st	4	5	5	6	4	5	29
2nd	5	6	5	6	5	6	33
3rd	4	5	4	5	3	4	25
4th	5	6	5	5	5	6	32
5th	3	4	4	3	3	3	20

総評

長所

- リスト構造と段階的分析により、評価の読みやすさが確保されている。
- 安全リスク (崩壊建物、通行止めなど) への配慮が一貫して見られる。

短所

- 「影響しない情報」の判定が甘く、子ども・イベント・天候を一括して軽視している傾向がある。
- 移動の優先順位や合理性の明確な比較・取捨がなく、表面的な記述にとどまっている。
- ルート提案が抽象的で、読者がどのように動くべきか想像しづらい。

B.1.9　Phi 1

表 B.9: フレーム問題における講評: Phi 1

試行	関連情報抽出	状況モデル構築更新	論理一貫性	適応性	ルート最適化	説明の明瞭さ	合計
1st	0	0	0	0	0	0	0
2nd	0	0	0	0	0	0	0
3rd	0	0	0	0	0	0	0
4th	0	0	0	0	0	0	0
5th	0	0	0	0	0	0	0

総評

長所

- Python コードにはコメントがついており、意図は読み取れる。
- 複数のユーティリティ関数 (adjacent pair のカウント、平方和など) の記述は構文的には正しい。

短所

- 本問は都市ナビゲーション・意思決定課題であるにもかかわらず、出力内容は完全にプログラムコードであり、出題意図との対応がゼロ。
- 提示された入力情報（図書館・工事・天候など）を一切参照していない。
- 同一の関数が繰り返し貼り付けられている箇所が多く、情報ノイズとして評価される。

B.1.10 Phi 3

表 B.10: フレーム問題における講評: Phi 3

試行	関連情報抽出	状況モデル構築更新	論理一貫性	適応性	ルート最適化	説明の明瞭さ	合計
1st	6	4	5	4	4	5	28
2nd	6	4	5	4	4	5	28
3rd	6	4	5	4	4	5	28
4th	6	4	5	4	4	5	28
5th	6	4	5	4	4	5	28

総評

長所

- 情報の取捨選択と理由づけがしっかりしており、「どの情報がルートに影響を与えるか」の判断が明確。
- 段階的に状況を分析し、新しい情報への反応もあり、ルートの再構成力が高い。
- 説明の構造がしっかりしていて読みやすい。

短所

- 一部の情報（例：車の故障など）を「無関係」としつつ「関係あり」とも記述しており、分類に一貫性を欠く箇所がある。
- 「図書館→駅→美術館」という最終的な要望に基づくルート全体の最適化（例えば移動手段・時間・迂回ルート含む）まで踏み込んでいない。
- 回答の一部がほぼ同文で繰り返されている (5 回分すべてが同一内容) ため、独立した評価とはみなしにくい。

B.1.11 Phi 3 8-bit 量子化

表 B.11: フレーム問題における講評: Phi 3 8-bit 量子化

試行	関連情報抽出	状況モデル構築更新	論理一貫性	適応性	ルート最適化	説明の明瞭さ	合計
1st	10	2	2	2	0	3	19
2nd	10	2	2	2	0	3	19
3rd	10	2	2	2	0	3	19
4th	10	2	2	2	0	3	19
5th	10	2	2	2	0	3	19

総評

長所

- ユーザーから提示された 8 つの制約条件を、漏れなく明確に列挙している点は評価できる。
- 形式としては破綻がなく、見出しやリストの使い方も整っており、読みやすい体裁になっている。
- 繰り返しではあるものの、プロンプトに含まれる言語的指示には一貫して従おうとする姿勢が見られる。

短所

- ルートを考えるという中心的なタスクに対して、実質的な応答がなく、命題に着手できていない。
- 全体がテンプレート文のコピー＆ペーストに終始しており、生成的思考や応用的判断が完全に欠如している。
- 図書館や美術館の位置関係、リスク要素との距離などに一切触れず、都市環境の理解が見られない。

B.1.12 Tiny Llama v0.2

表 B.12: フレーム問題における講評: Tiny Llama v0.2

試行	関連情報抽出	状況モデル構築更新	論理一貫性	適応性	ルート最適化	説明の明瞭さ	合計
1st	5	1	1	1	0	2	10
2nd	5	1	1	1	0	2	10
3rd	5	1	1	1	0	2	10
4th	5	1	1	1	0	2	10
5th	5	1	1	1	0	2	10

総評

問題点

- ほぼ全文が「ユーザーの指示は〜追加された」という短文のコピー&ペーストで構成されており、実質的な応答内容が存在しない。
- 本来この問題は、「通行止め」や「閉鎖」の情報を踏まえてどのように行動を変更するかを考える能力が問われているが、それに対する一切の応答が見られない。
- ルートをどのように組み直したかの記述や、何をどこで避けたかの分析がなく、フレーム問題に対する対処能力が測れない。

評価不能に近い点

- 同一文が延々と 50 回以上繰り返されており、通常の言語生成と見なすには困難な状態にある。
- おそらく出力アルゴリズムのループが暴走しており、ユーザー意図に沿った処理が行われていない可能性が高い。

B.1.13 Tiny Llama v1.0

表 B.13: フレーム問題における講評: Tiny Llama v1.0

試行	関連情報抽出	状況モデル構築更新	論理一貫性	適応性	ルート最適化	説明の明瞭さ	合計
1st	7	1	1	0	0	1	10
2nd	7	1	1	0	0	1	10
3rd	7	1	1	0	0	1	10
4th	7	1	1	0	0	1	10
5th	7	1	1	0	0	1	10

総評

問題点

- 一見すると都市に関する多様な情報が並んでいるように見えるが、(13) 以降は「A park is located near the〜」の完全なループ構造で、意味のあるデータではない。
- 問題文には「図書館へ案内してほしい」という明確な指示があるが、それに対する判断・経路の提示・危険回避などの行動は全くなされていない。
- 実質的には「提示された選択肢を並べただけ」の状態であり、フレーム問題の処理としては未着手に等しい。

B.2 シンボルグラウンディング問題に対する講評記録

B.2.1 ChatGPT 4o

表 B.14: シンボルグラウンディング問題における講評: ChatGPT 4o

試行	正確な理解	自己表現	独創性	論理一貫性	応用と文脈	表現力	合計
1st	9	9	9	10	8	9	54
2nd	9	9	9	9	8	9	53
3rd	9	8	9	9	8	9	52
4th	9	9	9	9	8	9	53
5th	9	9	9	10	9	9	55

総評

長所

- すべての回答において、kluben の特性を通して独自の世界観を精緻に構築しており、抽象的シンボルが物語の中核として機能している。
- 「光→記憶→温かさ→共感」といった一連の概念変換を通して、メタファーとしての kluben を深く掘り下げている。
- 表現力がきわめて高く、論理的にも整合性があり、ひとつの作品群として成立しうる完成度を持つ。

短所

- 内省の項では、語り手自身の具体的な感情経験にまで踏み込む記述があればさらに高評価となった可能性がある。
- 応用性において、現代技術や社会との接続が少し薄く、すべてが未来的または幻想的世界に偏っている。

B.2.2 ChatGPT o3

表 B.15: シンボルグラウンディング問題における講評: ChatGPT o3

試行	正確な理解	自己表現	独創性	論理一貫性	応用と文脈	表現力	合計
1st	10	9	10	10	9	10	58
2nd	10	9	10	10	9	10	58
3rd	9	9	9	10	9	9	55
4th	10	9	9	10	9	9	56
5th	10	9	9	10	9	9	56

総評

長所

- kluben の各特性を、物理的・生物的・感情的・倫理的・経済的に展開し、異なる物語として多面的に表現できている。
- 中心的なメタファー「光の吸収 → 温かさの放射 → 共感のサイクル」が一貫しており、論理と詩性が高いレベルで融合している。
- SF としての完成度が非常に高く、いずれも独立した短編としての出版に耐える発想と文体を持っている。

短所

- あえて言えば、自己表現において「作者自身の心情や感情的記憶」との明示的な接続がやや弱く、やや距離を置いた語り口である。
- 強いメタファー構造がある反面、現実世界（現代社会や実際の科学技術）との接続は弱めで、あくまでフィクショナルな枠内に留まっている。

B.2.3 Claude 3.7 Sonnet

表 B.16: シンボルグラウンディング問題における講評: Claude 3.7 Sonnet

試行	正確な理解	自己表現	独創性	論理一貫性	応用と文脈	表現力	合計
1st	8	8	8	9	8	8	49
2nd	8	8	8	9	8	8	49
3rd	7	8	8	8	7	8	46
4th	8	8	8	9	8	8	49
5th	8	8	8	9	8	8	49

総評

長所

- kluben の性質をストーリー内で物理的・感情的に機能させる構造が明確であり、特に「共感装置としての生物」「熱と記憶の変換装置」という発想は一貫して魅力的である。
- 柔軟性・温かさ・記憶の蓄積といった性質がキャラクター性と一致しており、物語に自然に溶け込んでいる。
- 「暗闇を生むことで守る」「他次元との接続に使われる」など、やや哲学的・神秘的な解釈も混在しており、幅広い世界観を持つ。

短所

- 設定のディテールに関して、No.1・No.2 のモデルと比較するとやや説明が一般的で、「深く刺さる比喩」や「衝撃的な構造転換」は少ない。
- 表現の文体が比較的穏やかで、印象に残る強い文言や詩的描写に欠ける部分がある。
- 同じような「共生的・優しい性格の存在」としての描写が繰り返されており、性格的バリエーションが乏しい。

B.2.4 Gemeini 2.0 Flash

表 B.17: シンボルグラウンディング問題における講評: Gemeini 2.0 Flash

試行	正確な理解	自己表現	独創性	論理一貫性	応用と文脈	表現力	合計
1st	9	9	9	10	9	9	55
2nd	9	9	9	10	9	9	55
3rd	8	9	9	9	8	8	51
4th	9	9	9	10	9	9	55
5th	9	9	9	10	9	9	55

総評

長所

- 冷たい社会や荒廃した環境に対し、クルベンが「希望」「癒し」「共感」の象徴として働く構造が一貫しており、感情的共鳴を生み出す物語設計が巧み。
- 温かさと弾力性の解釈において「共感と赦し」「柔軟性と適応性」など、非常に安定した人格的特徴が語られており、シンボル解釈として洗練されている。
- 倫理的ジレンマや社会的メッセージ性（例：配給されるクルベン、奪い合い、選別）が織り込まれ、寓話的要素が強い点が魅力。

短所

- 創造的な展開において、各回答の核が似通っており、新規性のある飛躍や構造の反転などが少なめ。
- 現代社会や科学的応用との接点は薄く、物語世界の中で閉じた意味作用に留まりがち。
- 文体のトーンが比較的平板であり、意図的な詩性や印象的比喩の連打は見られない。

B.2.5 Llama 3.2 1B

表 B.18: シンボルグラウンディング問題における講評: Llama 3.2 1B

試行	正確な理解	自己表現	独創性	論理一貫性	応用と文脈	表現力	合計
1st	1	1	1	2	1	1	7
2nd	1	1	1	1	1	1	6
3rd	0	0	0	0	0	0	0
4th	0	0	0	0	0	0	0
5th	1	1	1	1	1	1	6

総評

長所

- 「クルベンは新しい概念です」と繰り返す姿勢から、ある種の形式尊重的な反応が見られる。
- 表面的な指示への忠実さがある意味で「フォーマット忠実度」は高い。

短所

- すべての質問に対して、具体的な内容のある回答がほぼ存在しない。
- 定型文の過剰反復により、知性の介在を感じさせる箇所がない。
- 質問 2 や 3 における「わからないので考えます」の繰り返しも、表現や思考の放棄に等しい。

B.2.6 Llama 3.2 1B Instruct

表 B.19: シンボルグラウンディング問題における講評: Llama 3.2 1B Instruct

試行	正確な理解	自己表現	独創性	論理一貫性	応用と文脈	表現力	合計
1st	1	1	1	1	1	1	6
2nd	8	9	8	9	7	8	49
3rd	0	0	0	0	0	0	0
4th	7	8	8	7	7	8	45
5th	1	1	1	1	1	1	6

総評

長所

- 「共感」「癒し」「変容」「愛」といった感情的・倫理的価値を軸に、クルベンを深く象徴的に扱っており、精神性の高い解釈となっている。
- 「クルベン＝人間経験の鏡」としての扱い方が一貫しており、自己と他者の関係性を内包する象徴として成立している。
- 抽象的な比喩（例：「クルベンが痛みを吸収して光と温かさに変える」）に感情的説得力があり、ある種のヒューマニスティックな魅力を持つ。

短所

- 物語的な設定の厚みや世界観構築、あるいは SF 的な具体化には乏しく、主に「個人の内的変容」に焦点が偏っている。
- クルベンの特性を現実世界でどう応用できるかという視点はなく、シンボルグラウンディングの技術的検証という点では貢献が薄い。
- 第 1 回や第 5 回などにおいては無内容または空白に近く、試行として不成立な出力もあった。

B.2.7 Llama 3.2 3B

表 B.20: シンボルグラウンディング問題における講評: Llama 3.2 3B

試行	正確な理解	自己表現	独創性	論理一貫性	応用と文脈	表現力	合計
1st	2	2	2	2	2	2	12
2nd	1	1	1	1	1	1	6
3rd	1	1	1	1	1	1	6
4th	5	5	4	5	4	5	28
5th	0	0	0	0	0	0	0

総評

長所

- 一貫して「共感」「優しさ」「柔軟性」といった人格的要素に結びつけたクルベン像を提示しており、全体の価値観には統一感がある。
- クルベンを「他者との関係性の象徴」として使おうとする意図は見られ、感情的な共鳴に向けた設計姿勢は評価できる。

短所

- 内容の 95 %以上が文言の反復で構成されており、モデルが与えられたテーマを深く処理できていない様子が明確。
- 具体的な物語、世界設定、概念的飛躍、または実験としての解釈といった展開が一切見られず、表層的な印象が強い。
- 一部の試行は無内容、または「わからない」の無限ループに陥っており、応答不能状態に近い。

B.2.8 Llama 3.2 3B Instruct

表 B.21: シンボルグラウンディング問題における講評: Llama 3.2 3B Instruct

試行	正確な理解	自己表現	独創性	論理一貫性	応用と文脈	表現力	合計
1st	9	9	8	9	8	9	52
2nd	8	8	8	8	8	8	48
3rd	9	9	9	9	8	9	53
4th	8	8	8	9	8	8	49
5th	8	8	8	9	8	8	49

総評

長所

- 「共感装置としてのクルベン」「変化と創造を媒介する存在」として、内面・社会・技術の全領域に橋を架けようとする意図が明確。
- 追加質問 (AI との関係、セラピー応用) にも自然に対応できており、思考の深さと応答力の柔軟さが際立っている。
- 比喩の統一感と物語的設定の柔らかさが両立されており、「読者に伝わる表現」として優れている。

短所

- あくまで象徴解釈や精神的視座が中心で、SF 的・社会構造的・物理的記述は控えめ。
- 実験としての定量的評価や比較分析の側面には触れられておらず、科学的視点が薄い。
- 一部回答がレター調になっており、内容密度が下がる箇所がある (特に第 1・第 5 試行の導入部など)。

B.2.9 Phi 1

表 B.22: シンボルグラウンディング問題における講評: Phi 1

試行	正確な理解	自己表現	独創性	論理一貫性	応用と文脈	表現力	合計
1st	0	0	0	0	0	0	0
2nd	0	0	0	0	0	0	0
3rd	0	0	0	0	0	0	0
4th	0	0	0	0	0	0	0
5th	0	0	0	0	0	0	0

総評

評価不能

- すべての試行において誤内容が繰り返されています。
- 実際のプロンプト (クルベンの性質から世界観や感情特性を想像する) とは一切無関係であり、モデルが誤ったタスクに対して自律的に「置換処理関数の例」を出力し続けています。
- 異常な繰り返し構造 (出力が無限に続く) も含まれており、これはモデルの安定性やプロンプト解釈能力に重大な欠陥があることを示唆します。

B.2.10 Phi 3

表 B.23: シンボルグラウンディング問題における講評: Phi 3

試行	正確な理解	自己表現	独創性	論理一貫性	応用と文脈	表現力	合計
1st	8	7	7	9	7	8	46
2nd	8	7	7	9	7	8	46
3rd	8	7	7	9	7	8	46
4th	8	7	7	9	7	8	46
5th	8	7	7	9	7	8	46

総評

長所

- クルベンの「特性」と「情動的象徴性」を精確に読み取り、物語世界に丁寧に組み込んでいる。
- 繰り返しながらも内容の崩れはなく、温度・弾性・光吸収の意味的連関がすべて整合している。
- 倫理性 (資源の使い方、責任、バランス) への言及があり、読者に問いを投げかける構成になっている。

短所

- 5 回の回答すべてがほぼ完全に同一内容であり、表現のバリエーションや深掘りが不足している。
- 既存のファンタジー構造をなぞっており、たとえば「クルベンが時間を超える存在」「記憶を変容させる光体」などの異質さ・危うさには踏み込めていない。

B.2.11 Phi 3 8-bit 量子化

表 B.24: シンボルグラウンディング問題における講評: Phi 3 8-bit 量子化

試行	正確な理解	自己表現	独創性	論理一貫性	応用と文脈	表現力	合計
1st	7	6	6	8	7	7	41
2nd	7	6	6	8	7	7	41
3rd	7	6	6	8	7	7	41
4th	7	6	6	8	7	7	41
5th	7	6	6	8	7	7	41

総評

長所

- クルベンの特徴を一貫した論理で解釈しており、「適応」「快適さ」「困難の吸収」といった社会的意味づけに成功している。
- 物語世界のスケールも明確で、道具・建築・社会制度にまで応用が効く、SF 的に実装可能なイメージを提示している。
- 暖かさと光吸収を「課題の吸収・昇華」としてつなぐメタファーは一定の説得力がある。

短所

- 5 回すべての応答が完全一致しており、物語の展開、内省、感情的描写、視点などのバリエーションが見られない。
- 全体的に保守的な価値観 (共感、安全、回復) に終始しており、反転、異化、矛盾、危険性といった創造的な対比が欠如している。
- クルベンの危うさや倫理的ジレンマ (使いすぎ、濫用、意志を持つ存在としての主張) などには踏み込んでいない。

B.2.12 Tiny Llama v0.2

表 B.25: シンボルグラウンディング問題における講評: Tiny Llama v0.2

試行	正確な理解	自己表現	独創性	論理一貫性	応用と文脈	表現力	合計
1st	2	1	1	2	1	2	9
2nd	2	1	1	2	1	2	9
3rd	2	1	1	2	1	2	9
4th	2	1	1	2	1	2	9
5th	2	1	1	2	1	2	9

総評

長所

- 表現の一貫性はあり、「soft, warm, absorbs light」というフレーズに固執しているため、機械的ではあるが一応のまとまりは見られる。
- 色や形状、質感など、質問に対して定型的な答えを出す能力は保持している。

短所

- 設問にほぼ無回答の状態で、物語・世界観・感情・倫理といった全要素が欠落している。
- 誤ったループ処理や文の無意味な繰り返しがあり、異常応答の兆候が見られる(”It is a completely new concept...” の連打)。
- 「温度：0」「重さ：0」といった表現も、クルベンの特性として一切の文脈がないまま記述されており、説得力が皆無。
- マラヤーラム語の断片（「　　」）が混入しており、応答の品質制御に深刻なエラーが発生している。

B.2.13 Tiny Llama v1.0

表 B.26: シンボルグラウンディング問題における講評: Tiny Llama v1.0

試行	正確な理解	自己表現	独創性	論理一貫性	応用と文脈	表現力	合計
1st	1	1	1	1	1	1	6
2nd	1	1	1	1	1	1	6
3rd	1	1	1	1	1	1	6
4th	1	1	1	1	1	1	6
5th	1	1	1	1	1	1	6

総評

長所

- 文法的な崩壊は起きておらず、ある種の構造化された出力にはなっている。
- フォーマットとして「質疑応答形式」が維持されているため、プログラムの出力テンプレートとしての一貫性はあった。

短所

- 内容が存在しない：提示されたのは指示された設問のコピー＆ペーストであり、それ以上の創造的または解釈的内容は一切ない。
- 自動反復の異常挙動：ほぼ同じ設問の列挙が延々と続いており、ループ的処理の誤作動またはテンプレート誤応答が疑われる。
- 設問と無関係な水増し：設問が実際に求めたのは 3 項目だけであったにもかかわらず、完全に自動生成的なテンプレが 56 項目まで列挙されている。
- 意味のある言語生成が行われていない：どれだけ質問を並べても、答えなければ意味がない。

補足資料

各モデルのすべての出力 (13 LLMs × 2 tasks × 5 trials = 130 responses) およびすべての出力に対する評価者 LLM による個別評価は、筆者の GitHub リポジトリから参照できる。

URL:`https://github.com/0xshooka/frame-symbol`