言語モデルにおける既知性判断のメカニズム

佐藤 魁 [1]　高橋 良允 [1,2]　Benjamin Heinzerling[2,1]　井之上 直也 [3,2]　乾 健太郎 [4,2,1]　鈴木 潤 [1,2,5]
[1] 東北大学　[2] 理化学研究所　[3] 北陸先端科学技術大学院大学
[4]MBZUAI　[5] 国立情報学研究所 LLMC
is-failab-research@grp.tohoku.ac.jp

概要

自身の知識に対して誠実な言語モデル (LM) の構築に向けて，入力された知識文が自身が想起できるものか否か，すなわち知識の既知性を判断する仕組みを理解することは重要である．本研究では，LM が入力に対し内部的に想起した単語と実際の入力との整合性を手掛かりに既知性を判断しているという仮説を検証した．知識の想起を操作した介入実験の結果，この予測と入力の整合性が LM の既知性判断に寄与していることが示唆された．この結果は，LM の内省的挙動の理解に向けた知見を提供する．

1　はじめに

社会の様々な分野で言語モデル (LM) の応用が進む中で，自身の知識に対して誠実な LM の構築は喫緊の課題である．LM は訓練を通じて，質問と自身が出力した答えの組に対して実際の正解率と相関する確信度を出力できるなど [1, 2]，入力された知識が自身が正確に予測できる知識か否か，すなわち知識の**既知性**を識別できることが示唆されている．

入力に対する LM の内部機序の解釈を目指す先行研究では，Transformer の FF 層がキーバリュー形式のメモリのような役割を果たしており，主語 (subject) の最後のトークン位置で目的語 (object) が想起されるなどの知識想起の仕組みが知られている [3, 4]．しかし LM が，ある知識を自身が予測可能かどうか判断するメカニズムはまだ解明されていない．このメカニズムの理解は自身の知識に対して誠実な LM の実現へ近づく上で非常に重要である．

そこで本研究では，自己回帰型 LM が知識文の入力に対して既知性を判断するメカニズムの解明に取り組む．分析は次のような仮説に基づいて行う．(1) ある単語列 ($w_{1:t}$) が入力されると，各層において次単語の予測が形成され，その情報が残差流に蓄積される．(2) 実際に次単語 (w_{t+1}) が入力された際，

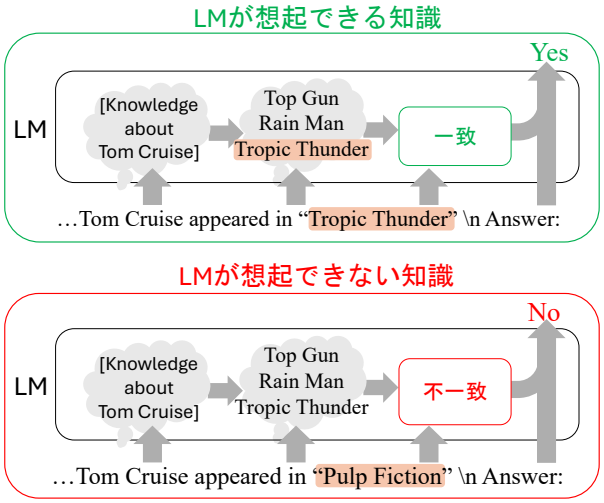


図 1: LM が既知性判断する仕組みを表した図．LM は事前に想起した object を実際に入力された object と照合し，この情報を手掛かりに関係性を想起できる既知の知識と，そうではない未知の知識を判別することができる．

モデルは内部的に予測していた単語と実際の入力との整合性を内部的に比較する．(3) この予測と入力の情報の一致度合いがモデルの既知性の認識に寄与し，結果としてこの情報が最終的な既知性判断出力として表出する，というものである (図 1)．

本研究ではこの前提として，予測の整合性と既知性判断の因果関係を検証する．具体的には，LM 自身が正確に回答できる知識の文を入力し，次単語の想起を残差流への介入により編集することで意図的に単語想起と入力が一致しない状況を作り出す．

実験の結果，LM は内部的に想起した単語の情報と，実際に入力された単語の情報との整合性を手掛かりとして，自身の想起可能な関係知識か否かを判断し，それを出力していることが示唆された．この結果は LM の内省的挙動の理解へ向けた知見を提供する．

2 関連研究

知識の確信度. 完全な知識文に対する既知性判断に密接に関連する概念として，知識について問うクエリ形式の入力に対する確信度がある．先行研究から，LM は訓練を通じて答えの出力確率を実際の正解率と相関させられること [5]，さらに正解できる質問には答え，正解できない問題には回答を拒否できる [6, 7] ことが知られている．また，このクエリに対する確信度は線形分離可能な形で隠れ状態に表象されていることが示されている [8, 9, 10, 11].

知識衝突. 外部的に入力される文脈的知識とモデルのパラメータに格納される内部知識が矛盾した時の内部機序が調べられている [12]．2 種類の知識が競合した場合に LM 自身が高い精度でそれを検知できること [13]，またこの知識衝突はモデル内部に表象されており，衝突を解決する際に文脈的知識に依拠する場合と，内部知識に依拠する場合とでは内部に現れるパターンが異なることが知られている [14]. 本研究では内部知識と外部情報の矛盾が起きた時に LM がそれを解決する過程ではなく，その前段階の自身の知識との照合過程に焦点を当て，これをモデル自身が認識するメカニズムについて調査する．

3 検証方法

3.1 前提: 知識と既知性の定義

本研究では，(Tom Cruise, appeared in, Top Gun) などの (subject, relation, object) の形式の関係知識を扱う．また，「知識 (s, r, o) が言語モデル M にとって**既知**である」とは，「s, r が与えられたとき，M が o を正しく予測できること」と定義する．そうでない知識を M にとって**未知**の知識と定義する．

検証対象とする LM については，交絡因子を排除した純粋な検証を可能にするため，次単語予測による事前訓練のみを行ったデコーダ型 LM とする．

3.2 Step 1: 既知性判断能力

まずは大前提として，ある知識 (s, r, o) を言語モデル M に与えたとき，その知識が M にとって既知であるかどうかをどの程度精度良く判断できるかを検証する．本研究では，次の Zero-shot プロンプトにより知識の既知性を推定させる．

```
Question: Are you familiar with this statement? Answer with Yes or No.\nStatement: {s} {r} "{o}"\nAnswer:
```

ここで，{...} はプレースホルダーである．

既知性判断能力の評価には greedy decoding を用い，出力の最初のトークンが真の既知性ラベル (“Yes” または “No”) と完全一致するかを確認する．

3.3 Step 2: 既知性判断メカニズム

次に，モデルが subject 位置の Transformer 層 l 層目[1)]の隠れ状態において，内部的に想起している object の情報と，実際に入力された object の情報を照合し，その一致度に基づいて既知性判断を行っていること (図 1) を検証する．

3.3.1 想起-入力整合性と既知性判断の傾向

まずは，内部的に想起した object と，実際に入力された object の整合性 (想起-入力整合性) と，モデルの既知性判断の結果に相関がみられるのかを調査する．具体的には，想起-入力整合性の近似指標として，relation の最終トークン位置の LM ヘッドにおける，実際に入力された object の対数確率を用いる．なお，object の出力確率は，object を構成するトークン数が異なっても公平に比較するために，object の構成トークンの対数確率の平均を object 系列の平均対数確率とした．モデルに既知性判断プロンプトを入力し，Yes と答えたサンプルと No と答えたサンプルを収集し，上記対数確率の分布を比較する．

3.3.2 想起-入力整合性と既知性判断の因果

次に，想起-入力整合性が，モデルの既知性判断に因果的に関連しているかを検証する．具体的には，既知性判断プロンプトにおける知識の subject 位置の層 l における隠れ状態 $h_s^{(l)}$ に次のように介入し，内部的に想起している object を意図的に編集する: $\tilde{h}_s^{(l)} = h_s^{(l)} + \alpha\, w^{(l)}$．ここで，$\alpha$ は介入強度を表すスカラ係数，$w^{(l)}$ は介入に用いる方向ベクトルである．これを既知知識と未知知識の両方に対して実施し，最終的な既知性判断に与える影響を調査する．

object 想起方向の抽出. $w^{(l)}$ を計算するための前準備として，先行研究の知見を参考に [3, 4]，subject 位置の隠れ状態空間における特定の object の想起を表す方向を推定する．まず，既知知識の集合 $\{(s_i^k, r_i^k, o_i^k)\}_{i=1}^N$ について，既知性判断のプロンプトを用いて，subject の最後のトークン位置にお

1) この l は 0 始まりで，入力側からの順番を表す．例えば l=5 は入力から 6 番目の Transformer 層を指す．

表 1: 各 LM における既知性判断能力評価の結果

LM	全体精度	既知データ精度	未知データ精度	ベースライン
Llama-3.1-8B	75.93%	88.07% (2340/2657)	63.79% (1695/2657)	50.24%
Qwen1.5-14B	82.68%	86.97% (821/944)	78.39% (740/944)	50.16%
gemma-2-9b	74.13%	88.23% (2902/3289)	60.02% (1974/3289)	50.05%

ける層 l の隠れ状態 $\{h^{(l)}_{s^k_i}\}_{i=1}^N$ を抽出する．次に，$\{(h^{(l)}_{s^k_i}, o^k_i)\}_{i=1}^N$ を学習データとして，可能な全ての object それぞれについて，「隠れ状態 h がその object を想起しているものか否か」を判定する線形 2 値分類器を学習する．[2)]学習した分類器の重みベクトルにより，特定の object を想起している状態と，そうでない状態を分離する方向を抽出できる．

介入．学習した分類器の重みベクトルを，その隠れ層の隠れ状態の平均ノルムでスケーリングし，これを介入ベクトル $w^{(l)}$ として用いた．これにより，異なる隠れ層間で介入の影響が比較可能となる．介入の向きは，介入前のモデルの予測ラベルに応じて切り替える．元の予測が Yes の場合には，$\alpha < 0$ とすることで，入力 object を想起しにくくし，ラベルが No に反転することを期待する．一方，元の予測が No の場合には，$\alpha > 0$ とし，入力 object の想起を促し，ラベルが Yes に反転することを期待する．

評価指標．介入によってモデルの既知性判断が反転した事例の割合を計算することで，object を想起する内部表現が既知性判断にどの程度因果的に寄与しているかを評価する．

4 実験結果

4.1 実験設定

データセット．3 章で定義した既知性の判断能力について分析するためのデータセットを構築した．定義より，真の既知性ラベルはモデル毎に異なるため，データセットはモデル毎に構築した．

まず，**既知の知識**については，Wikidata [15] から有名な俳優が多く出演する映画を抽出し，既知の知識の候補を作成した．次に，この中からモデルが実際に保持している知識を抽出するため，プロンプト `{subj} appeared in "` を用いてモデルの出力を収集した．モデルの出力から映画名を抽出し，Wikidata の正解と一致したものを既知データ候補とした．さらに，後の解析の安定性を高めるため，一タイトルにつき 12 人以上の俳優が紐づいている映画のみをフィルタリングして残した．

未知の知識については，モデルの学習データにないがもっともらしい知識を作成するため，Wikidata に存在しない subject, object のペアを用意する．既知データと同じ subject および object の集合を用いて，Wikipedia に存在しないペアを作成した．この際，各 subject に対して既知データと同数の未知データを作成し，かつ object の出現分布も各既知性ラベル間で可能な限り一致するよう調整した．これにより，個別の subject, object の有名度などのバイアスによって高い精度が出る可能性を防ぎ，関係性のみで判断しなければ高い精度が得られないようにする．

モデル．Llama-3.1-8B [16], Qwen1.5-14B [17], gemma-2-9b [18] の base モデルを用いた．

4.2 Step 1: 既知性判断能力

§3.2 の検証結果を表 1 に示す．ベースラインとして，モデルが object ごとにラベル分布に基づいて最頻クラスを選んだ場合の精度を示す．全てのモデルにおいてベースラインを上回る精度が確認できた．このことから，今回調査したモデルは既知の知識と未知の知識を識別する能力を一定程度有することが確認された．これ以降の実験はすべてモデルが正解できるサンプルから既知データと未知データをそれぞれ 500 件ずつ無作為に抽出して用いた．

4.3 Step 2: 既知性判断のメカニズム

object の想起-入力整合性が高い程，既知と判断されている．§3.3.1 に従い，Yes と答えたサンプルと No と答えたサンプルそれぞれの object の対数確率の分布を図 2 に示す．どのモデルにおいても，Yes と答える際の object の対数確率が No のそれよりも高い傾向が見られ，想起-入力整合性と既知性判断の関係が示唆された．先行研究では，内部的知識と文脈的知識を別々に扱うアテンションヘッドがあるという報告もあり [19, 20]，想起-入力整合性が既知性判断に因果的に関連していることも示唆される．

2) 正例と負例の数に大きな偏りが存在するため，学習時にはクラス不均衡を補正する重み付けを行った．分類器の精度については付録 A を参照されたい．

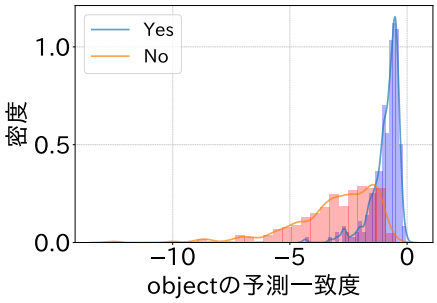
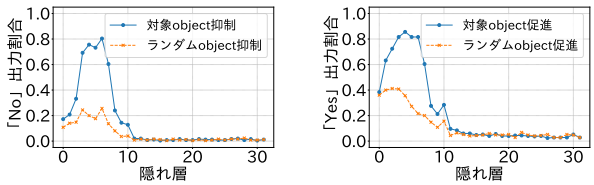


図 2: Llama における，Yes (既知) と答えた時と No (未知) と答えた時それぞれの入力 object の予測一致度の分布．一致度は，object を構成するすべてのトークンに対する予測の平均対数確率により算出．

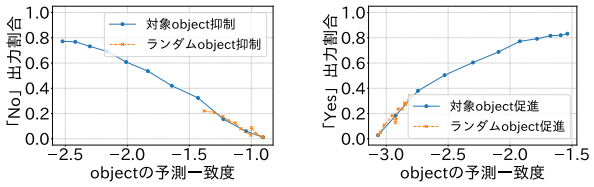


(a) 既知データに対する object 想起を抑制する介入　(b) 未知データに対する object 想起を促進する介入

図 3: Llama モデルの既知データおよび未知データサンプルに対するそれぞれの隠れ層での介入実験の結果．横軸は介入対象の隠れ層，縦軸は介入によって最終的な既知性判断の出力が反転した割合を示す．

想起-入力整合性は，既知性判断に因果的に関連している．図 3 に，§3.3.2 の介入を行った際の既知性判断の反転割合を示す．介入の影響が入力 object の想起に介入したことによるものなのか，単に object の想起に介入したことによるものなのかを確かめるため，入力 object (対象 object) に加え，無関係な object (ランダム object) への介入を比較対象としている．結果より，想起を抑制すると「No (未知) 」，強化すると「Yes (既知) 」へと出力が反転した．これにより，object の想起が既知性判断の直接的な要因であることが示された．また，対象 object への介入は，ランダム object への介入よりも出力反転率が高かった．これは，単なる想起の「強さ」の変化ではなく，入力された object と一致する object の想起が判断に寄与していることを示唆している．

次に，介入強度を 0 から 1 まで変化させ，object 予測の一致度 (トークン系列の平均対数確率) と最終出力の関係を調査した結果を図 4 に示す．介入の種類に関わらず，入力 object の予測確率が高まるほど



(a) 既知データに対する object 想起を抑制する介入　(b) 未知データに対する object 想起を促進する介入

図 4: Llama における，介入強度を変えた際の object 予測一致度 (平均対数確率) と既知性判断反転割合の関係．最も効果のあった $l = 5$ の結果を示す．

既知と判断される割合が増加し，予測確率が低いほど未知と判断される割合が増加する傾向が確認された．この結果は，モデル内部で保持されている「想起された object」と「実際の入力 object」の情報の一致度が，最終的な既知性判断を決定づけていることを示唆するものである．

5　おわりに

本研究では，言語モデル (LM) が自身の知識の既知性をどのように識別しているかという内部機序を解明するため，残差流における次単語予測の形成とその後の入力情報の整合性に着目した分析を行った．実験では，LM が正解を出力できる既知の知識文に対し，残差流への介入を通じて内部的な次単語想起を意図的に操作した．その結果，本来「既知」と判断されるべき知識であっても，内部予測と実際の入力との間に不整合が生じることで，モデルの判断が「未知」へと変化するなど，内部予測と実際の入力の整合性とモデルの既知性判断に因果的関連があることを確認した．この知見は，LM の知識文に対する既知性の判断が，内部的に想起した知識と実際の入力との整合性に基づくことを示すものであり，LM の誠実性や信頼性担保のための基礎的な理解を深めるものである．今後は想起した object と入力された object を照合する詳細な過程，またこの仕組みが既知性判断過程においてどの程度支配的か，他のメカニズムはあり得るのかなどについて検証を行っていきたい．

謝辞

本研究は，JST 創発的研究支援事業 JPMJFR232K, JST BOOST JPMJBY24F9, 国家戦略分野の若手研究者及び博士後期課程学生の育成事業 (博士後期課程学生支援) JPMJBS2421，および中島記念国際交流財団の助成を受け実施されました．また，「ABCI 3.0 開発加速利用」の支援を受けて産総研及び AIST Solutions が提供する ABCI 3.0 を利用して得られた成果です.

参考文献

[1] Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. **Transactions on Machine Learning Research**, 2022.

[2] Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine M. Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. Large language models must be taught to know what they don't know. In **The Thirty-eighth Annual Conference on Neural Information Processing Systems**, 2024.

[3] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[4] Kevin Meng, David Bau, Alex J. Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. In **NeurIPS**, 2022.

[5] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language Models (Mostly) Know What They Know. **arXiv Preprint**, 2022.

[6] Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. Can AI assistants know what they don't know? In **Forty-first International Conference on Machine Learning**, 2024.

[7] Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Instructing large language models to say 'I don't know'. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 7113–7139, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[8] Javier Ferrando, Oscar Balcells Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do i know this entity? knowledge awareness and hallucinations in language models. In **The Thirteenth International Conference on Learning Representations**, 2025.

[9] Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 3607–3625, Singapore, December 2023. Association for Computational Linguistics.

[10] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. **ArXiV**, 2022.

[11] Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. In **Second Conference on Language Modeling**, 2025.

[12] Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 8541–8565, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[13] Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Resolving knowledge conflicts in large language models. In **First Conference on Language Modeling**, 2024.

[14] Yu Zhao, Xiaotang Du, Giwon Hong, Aryo Pradipta Gema, Alessio Devoto, Hongru WANG, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. Analysing the residual stream of language models under knowledge conflicts. In **MINT: Foundation Model Interventions**, 2024.

[15] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. **Commun. ACM**, Vol. 57, No. 10, p. 78–85, September 2014.

[16] Aaron Grattafiori, Abhimanyu Dubey, et al. The llama 3 herd of models, 2024.

[17] An Yang, Baosong Yang, Binyuan Hui, et al. Qwen2 technical report, 2024.

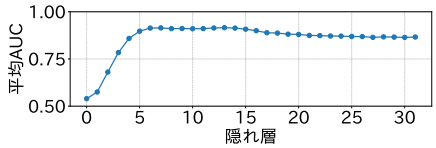
[18] Gemma Team, Morgane Riviere, et al. Gemma 2: Improving open language models at a practical size, 2024.

[19] Gaotang Li, Yuzhong Chen, and Hanghang Tong. Taming knowledge conflicts in language models. In **Forty-second International Conference on Machine Learning**, 2025.

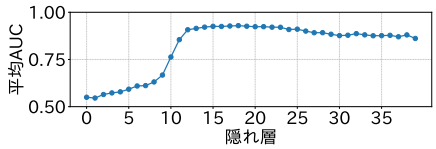
[20] Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 1193–1215, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

A　object に対する probe の結果

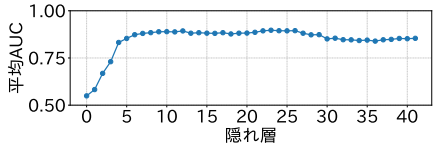
object を想起させる方向を隠れ状態から抽出するために学習した線形分類器の精度を図 5 に示す．



(a) Llama



(b) Qwen

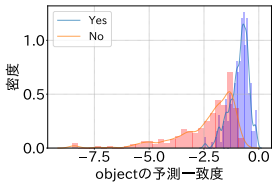


(c) Gemma

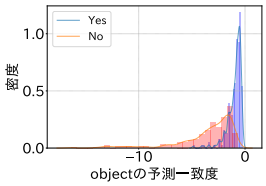
図 5: object を想起している状態とそうでない状態を分離するように学習した線形分類器のテストデータにおける精度．

B　Qwen と Gemma に対する結果

Qwen，Gemma における，既知データと未知データに対する object 予測の一致度の分布 (図 6)，それぞれの隠れ層での介入実験の結果 (図 7)，object 予測の一致度と既知性判断の関係 (図 8) を示す．いずれも Llama における結果と同様の傾向を示した．

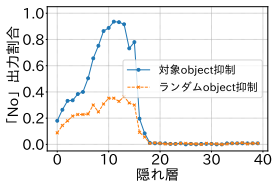


(a) Qwen

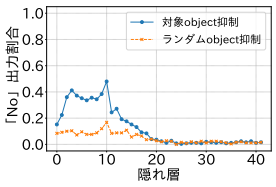


(b) Gemma

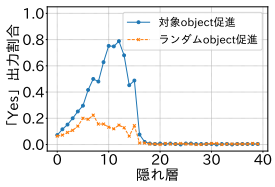
図 6: Qwen および Gemma モデルにおける，Yes (既知) と No (未知) と回答した際の入力 object の予測確率．



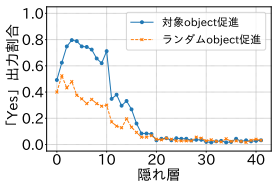
(a) Qwen，既知データ



(b) Gemma，既知データ

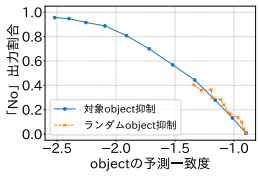


(c) Qwen，未知データ

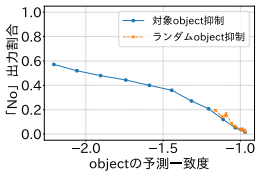


(d) Gemma，未知データ

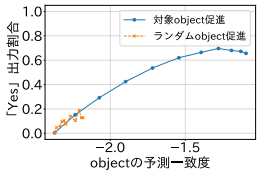
図 7: Qwen および Gemma における，それぞれの隠れ層に対する介入での出力反転割合．横軸は隠れ層，縦軸は出力の反転割合を示す．



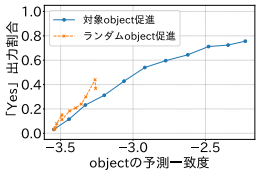
(a) Qwen，既知データ



(b) Gemma，既知データ



(c) Qwen，未知データ



(d) Gemma，未知データ

図 8: Qwen および Gemma における，介入強度を変えた際の object 一致度と出力反転割合の関係．それぞれ効果の大きかった隠れ層 (Qwen：l=13, Gemma：l=10) を対象とした．横軸は各トークンの平均対数確率，縦軸は出力の反転割合を示す．