論文紹介

Beyond Accuracy: Evaluating the Reasoning Behavior of Large Language Models - A Survey

紹介者：品川 政太朗

※以下、登場する図は論文からの引用です

書誌情報: Beyond Accuracy: Evaluating the Reasoning Behavior of Large Language Models - A Survey (COLM2024)

https://openreview.net/forum?id=Lmjgl2n11u#discussion

著者：Philipp Mondorf, Barbara Plank
MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
Munich Center for Machine Learning (MCML), Munich, Germany

LLMのReasoning Behaviorに注目したサーベイ論文

- タスク成功率である「{task, reasoning} performance」と対比された概念
- **スコアが上がればなんでもいいという立場ではなく、推論時の振る舞いに焦点を当てている**

> **Definition 2.2** (Reasoning Behavior). *The system's computed response to a reasoning task (the stimulus), particularly its actions, expressions and underlying mechanisms exhibited during the reasoning process.*

※以下、本紹介で「推論」は「reasoning」の意味で呼びます

論文の選定理由：

- Reasoningの隆盛を感じており、キャッチアップするため

タスク成功率が上がれば「推論能力が向上してる」は正しいの？

難しいタスクが解けているのは推論能力の向上ではなく訓練データの記憶に由来している可能性がある (Wu et al., 2024; Dziri et al., 2023; Razeghi et al., 2022; Zhang et al., 2023)

Reasoning Behaviorに焦点を当てる

- タスク成功率だけではなく、**推論の過程におけるLLMの振る舞い**にも注目
- 推論能力の検査用タスク（Reasoning Tasks）を設計しLLMの振る舞いを観察

本研究の問い：

- RQ1: 多様な推論タスクに対してLLMがどう振舞うのか？（前半の話）
- RQ2: LLMにおけるReasoning Behaviorを評価する方法のトレンドは何か？（後半の話）

COLM査読者の反応集　だいたいが好意的で全員6: Marginally above acceptance threshold以上

評価されているポイント

- トピックの重要性と網羅性
- evaluation typeの類型化など、LLMの推論能力について大きな視点でまとめている

懸念を示されているポイント

- CoTのようなpromptingアプローチによるLLM推論の改善には触れていない
- 主張の不正確さ（特にbrittlenessとmemorizationにより何が示唆されるのか）
- 要点と本論における重要な知見の強調が不足している（今後必要な研究について議論不足など）

その他に興味深い点

- Reviewer bLrd「LLMからReasoning Behaviorを引き出すにはどのようにしたらよいかという議論が欲しい」👈わかる

用語の定義: ReasoningとReasoning Behaviorについて

Reasoningは昔から色々な議論があるので本論文では改めて以下のように定義

> **Definition 2.1** (Reasoning). *The process of drawing conclusions based on available information (usually a set of premises).*

前提となる情報から結論を導き出す過程

Reasoning Behaviorは行動心理学の考え方を参考に、以下のように定義

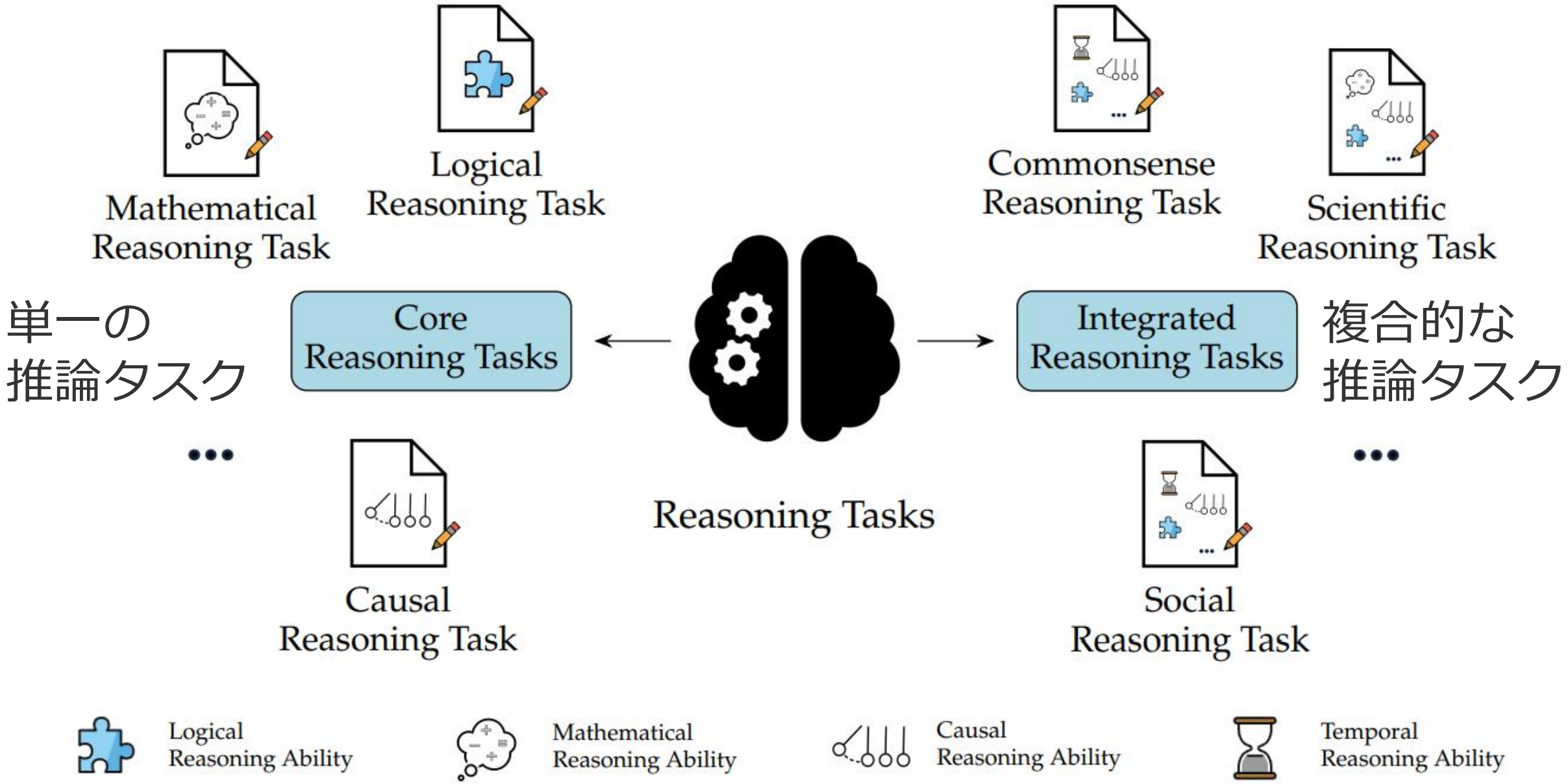
> **Definition 2.2** (Reasoning Behavior). *The system's computed response to a reasoning task (the stimulus), particularly its actions, expressions and underlying mechanisms exhibited during the reasoning process.*

reasoning taskを刺激と見立てて、その反応を評価する
- LLMの行動や表現、その土台となるメカニズムなど

前半の話
RQ1: 多様な推論タスクに対してLLMがどう振舞うのか？

- 得られた知見や傾向のまとめ
- LLMは結局訓練データに類似するパターンには強いが、分布外の設定には弱い

Reasoning Tasksの2類型：{Core, Integrated} Reasoning Tasks



(Core) Logical Reasoning Tasks

論理的なルールのもとでの前提から結論を導き出すタスク（例：三段論法）

Fact1: Aristotle is a human Rule: All humans will die Fact2: Aristotle will die		
Deduction (Fact1 + Rule → Fact2)	**Abduction** (Fact1 + Rule ← Fact2)	**Induction** (Fact1 + Fact2 → Rule)

"Fact" denotes specific knowledge while "rule" denotes general principle.

演繹的推論
前提となる知識・法則から結論を導く

アブダクション
事実と結論からその間の現象を推論する

帰納的推論
知識から一般化された法則を推論する

必ず結論がある

仮説を立てる（仮説が正しいとは限らない）
タスクはこの２つのどちらか

図は下記より引用
Natural Language Reasoning, A Survey https://dl.acm.org/doi/10.1145/3664194

(Core) Logical Reasoning Tasks

論理的なルールのもとでの前提から結

知識：アリストテレスは人である

法則：人はみな死ぬ

結論：よって、アリストテレスは死ぬ

Fact1: Aristotle is a human
Rule: All humans will die
Fact2: Aristotle will die

Deduction	**Abduction**	**Induction**
(Fact1 + Rule → Fact2)	(Fact1 + Rule ← Fact2)	(Fact1 + Fact2 → Rule)

"Fact" denotes specific knowledge while "rule" denotes general principle.

演繹的推論
前提となる知識・法則から結論を導く

アブダクション
事実と結論からその間の現象を推論する

帰納的推論
知識から一般化された法則を推論する

図は下記より引用
Natural Language Reasoning, A Survey https://dl.acm.org/doi/10.1145/3664194

(Core) Logical Reasoning Tasks

論理的なルールのもとでの前提から

知識：アリストテレスは人である

知識：アリストテレスはいずれ死ぬ

法則：死ぬ原因は全ての人が死ぬ法則が適用されたためだと考えられる

Fact1: Aristotle is a human
Rule: All humans will die
Fact2: Aristotle will die

Deduction	**Abduction**	**Induction**
(Fact1 + Rule → Fact2)	(Fact1 + Rule ← Fact2)	(Fact1 + Fact2 → Rule)

"Fact" denotes specific knowledge while "rule" denotes general principle.

演繹的推論
前提となる知識・法則から結論を導く

アブダクション
事実と結論からその間の現象を推論する

帰納的推論
知識から一般化された法則を推論する

図は下記より引用
Natural Language Reasoning, A Survey https://dl.acm.org/doi/10.1145/3664194

(Core) Logical Reasoning Tasks

論理的なルールのもとでの前提から結

知識：アリストテレスは人である

知識：アリストテレスはいずれ死ぬ

法則：すべての人がいずれ死ぬかもしれない

Fact1: Aristotle is a human Rule: All humans will die Fact2: Aristotle will die		
Deduction (Fact1 + Rule → Fact2)	**Abduction** (Fact1 + Rule ← Fact2)	**Induction** (Fact1 + Fact2 → Rule)

"Fact" denotes specific knowledge while "rule" denotes general principle.

演繹的推論

前提となる知識・法則から結論を導く

アブダクション

事実と結論からその間の現象を推論する

帰納的推論

知識から一般化された法則を推論する

図は下記より引用

Natural Language Reasoning, A Survey https://dl.acm.org/doi/10.1145/3664194

(Core) Deductive Reasoning TasksでのLLMの振る舞い

表面的なパターンマッチに引っ張られている恐れがある

- 大きいLLM+CoTは前提に即して単一のルールから結論を導ける（validityとatomicityが高い）
  - が、間違えると回復が難しい（utilityが低い）

生成された推論ステップを一階述語論理に変換してvalidity, atomicity, utilityを評価
Saparov & He (2023)

- Einsteinのパズルでは類似のパターンで正解率が高い
  - が、新しい問題には対応できなかった
  - 自己回帰モデルでは初期のエラーが後段の推論に大きな影響を与える

推論過程を計算グラフにパースしてLLMの多段階推論を評価
Dziri et al. (2023)

- 与える前提の順序をランダムに配置すると正解率が低下
  - 訓練データの頻出パターンに依存している疑い

Chen et al. (2024b)

- LLMは論理的否定の解釈が苦手
  - GPT-4がDe Morganの法則を正確に理解できてない

Sanyal et al. (2022)
Truong et al. (2023)

LLMは人間と同じように認知バイアスに影響される

三段論法における人間の論理的誤謬と同様のバイアスがLLMsにも見られる

- Eisape et al. (2024)

LLMは人間と同様に問題の意味的内容に影響される

- Dasgupta et al. (2022)

モデルの内部の挙動から考察する試み（Mechanistic Evaluation）

Hou et al. (2023)
- GPT-2やLLaMAモデルの注意パターンを分析
- マルチステップの推論プロセスが段階的に進行することが注意から分かる
- 層ごとにprobingすると
  - 低層ではタスクに関連する情報が出てくる
  - 高層では複雑な推論が行われている

Pirozelli et al. (2023)
- RoBERTa-large modelのprobingでも同様
- 上位層が推論に重要

Facts: St Johnstone is a Scottish team. St Johnstone is part of the Scottish Premiership.
Rules: If a team is part of the league, it has joined the league. St Johnstone and Minsk are different teams. For two different teams, either one team wins or the other team wins. Minsk won against St Johnstone.
Hypothesis: At least one Scottish team has joined the Scottish Premiership.
Label: TRUE

知識の組み合わせの２値分類

Pirozelli et al. (2023)

Dutta et al. (2024)
- LLaMA 2-7BをCoT promptingした時の内部状態の解析
  - 低層：トークン表現は事前学習で得られた分布に偏っている
  - 高層：トークン表現はコンテキスト内の事前分布に急激にシフトしている

(Core) Inductive Reasoning Tasks

Deductive Reasoningと比べるとあまりやられてないし、難しい
本質的なパターンの抽出ができておらず、冗長な情報を含みがち

- 事実から一般的なルールを導くこと自体は可能
  - ただし、正しいとは言ってない（前提の事実と一致しない、現実の知識に合致しない、冗長がち）

Yang et al. (2024)
GPT-JやLLaMA 7Bなどのモデルが与えられた事実から一般的なルールを導き出す能力を調べた

- モデルはルールを生成できる
  - ルールの適用においてしばしばエラーを犯し、人間が導き出すルールとは異なる傾向あり
  - タスクの記述が少し変更されただけでモデルの推論が著しく変わる傾向

Qiu et al. (2024)
GPT-3.5やGPT-4、Claude 2などのモデルがルールを導き出すとともに、それを適切に適用する能力を評価

- GPT-4は人間の判断に近い振る舞いを示す
  - 非単調性（non-monotonicity）をうまく扱えない：前提に追加の情報が与えられることで尤度が下がるケース
    例：{crow, peacock, **rabbit**} → bird （rabbitが加わると尤度が下がってほしいがLLMは下がらない）

Han et al. (2024)
GPT-3.5やGPT-4が特定のカテゴリー間の共通属性を推論するタスクで評価

(Core) Abductive Reasoning Tasks

Inductive Reasoningと同様、ハルシネーションが起きやすい
※使っているモデルがGPT3.5だったりするので注意

- GPT-3では事実に対して可能性のある説明をする能力は限定的
- 訓練データに含まれない予測や想像力を要する場面では人間が優れている

Collins et al. (2022)
GPT-3を使って、ある状況において予測される結果が現実と異なる場合にその理由を説明するタスクを評価

- LLMはしばしば矛盾する説明を生成し、同じ根拠で仮説を強化および弱化するなど、一貫性に欠ける回答を出す傾向

Rudinger et al. (2020)
GPT-2やBART、T5などのモデルがある仮説に対する根拠を強化または弱化する能力を

- LLMはハルシネーションしがち
- アブダクションタスクでは多段階推論能力が必要

Xu et al. (2023)
GPT-3.5やChatGPT、PaLM 2を対象にアブダクションタスクでの推論過程とエラーの傾向を調査

(Core) Mathematical Reasoning Tasks

数学の問題、計算タスクでも一貫性のなさがある

- 異なる問題表現に対してモデルの一貫したパフォーマンスが見られず、記憶に依存している傾向

Srivastava et al. (2024)
MATHデータセットを使用し、数学的な問題解決能力を評価、異なる表現で問題が構成されている

- 問題の構成や数値が変わるとモデルの正確さが低下する
- 訓練データに頻出しない数値や数式の形式が出題されると、モデルのパフォーマンスが大幅に下がる傾向

Razeghi et al. (2022)

- GPT-3.5などのモデルが文章問題において問題解決に無関係な情報が含まれる場合に混乱しやすい
- 無関係な情報が元の問題と類似の語彙や構成を持つと、さらに混乱が生じやすい

Shi et al. (2023)

(Core) Mathematical Reasoning TasksにおけるLLMの人間的なバイアス

人間が誤りやすいところではLLMも気を付けて推論する
難しい問題を簡単なタスクに置き換えて解こうとする傾向がある

Hagendorff et al. (2023)

- 人間が誤りやすい直観的問題（例えば、認知反射テストのような問題）で、より慎重な判断を行う能力が確認
- GPT-3が直観に基づく誤った解答を提供する一方で、GPT-3.5やGPT-4はより熟慮的な解答を示し、直観的な誤答を回避する傾向

McKenzie et al. (2023)
四捨五入タスク

- モデルが小数点以下の桁数に基づいて誤った四捨五入をする傾向
- 難しい問題を簡単なタスクに置き換えて処理する「属性代替（attribute substitution）」と呼ばれる人間の認知バイアスと類似

(Core) Mathematical Reasoning Tasksにおける内部状態の調査

層ごとの役割の違いがみてとれるらしい

- LLaMAの各層を分析し、上位層の方が数学的な問題解決に優れている一方、下位層では基本的な計算や知識が不足している傾向
- モデル内の層ごとに異なる役割がありそう

Chen et al. (2024a)

(Core) Causal Reasoning Tasks

新しい状況における因果関係の構築や反事実的なシナリオの理解に課題
反事実的な設定においては本質的な理解が難しく、単なる関連性の参照に留まりがち

- GPT-3やOPT、AlephAlphaのLuminousなどは訓練データ内で見られる因果関係については適切に回答できる
- が、新しい因果関係の構築が難しい

Zečević et al. (2023)

- LLMsは関連性のある質問には比較的正確に回答できる
- 介入や反事実のレベルの推論には苦戦する

Jin et al. (2023)
「因果のはしご（Ladder of Causation）」に基づき、3つの因果レベル（関連、介入、反事実）でモデルを評価

- 多くのLLMがデータの分布外のシナリオにおいて誤った推論を行いやすい
- ファインチューニングによって性能は向上するものの、未見のデータには対応が難しい

Jin et al. (2024)
変数間の相関関係から因果関係を推測するタスクでモデルを評価

(Core) Causal Reasoning Tasks

- 候補の因果関係が提示されている場合にはモデルが正確に推論できる
- 候補がない場合は因果関係を正確に見抜くことが難しい

Kosoy et al. (2023)
「ブリケット検出タスク」どのオブジェクトが光を点灯させる原因であるかを判断

- GPT-3などのモデルが反事実的な仮定のもとで結果を予測する能力が人間に比べて著しく低い
- 文脈における単純なキューに頼りがちで、反事実の本質的な理解が欠如している

Frohberg & Binder (2022)
Li et al. (2023)
反事実推論タスク

- 反事実を含む質問に対して、GPT-3などの「クローズドブック」モデルが誤った事実や不正確な前提に基づく回答を出す傾向

Yu et al. (2023)
反事実推論タスク

前半のまとめ

RQ1: 多様な推論タスクに対してLLMがどう振舞うのか？

結論：LLMは結局訓練データに類似するパターンには強いが、分布外の設定には弱い（確率的オウムの域を出ていない）

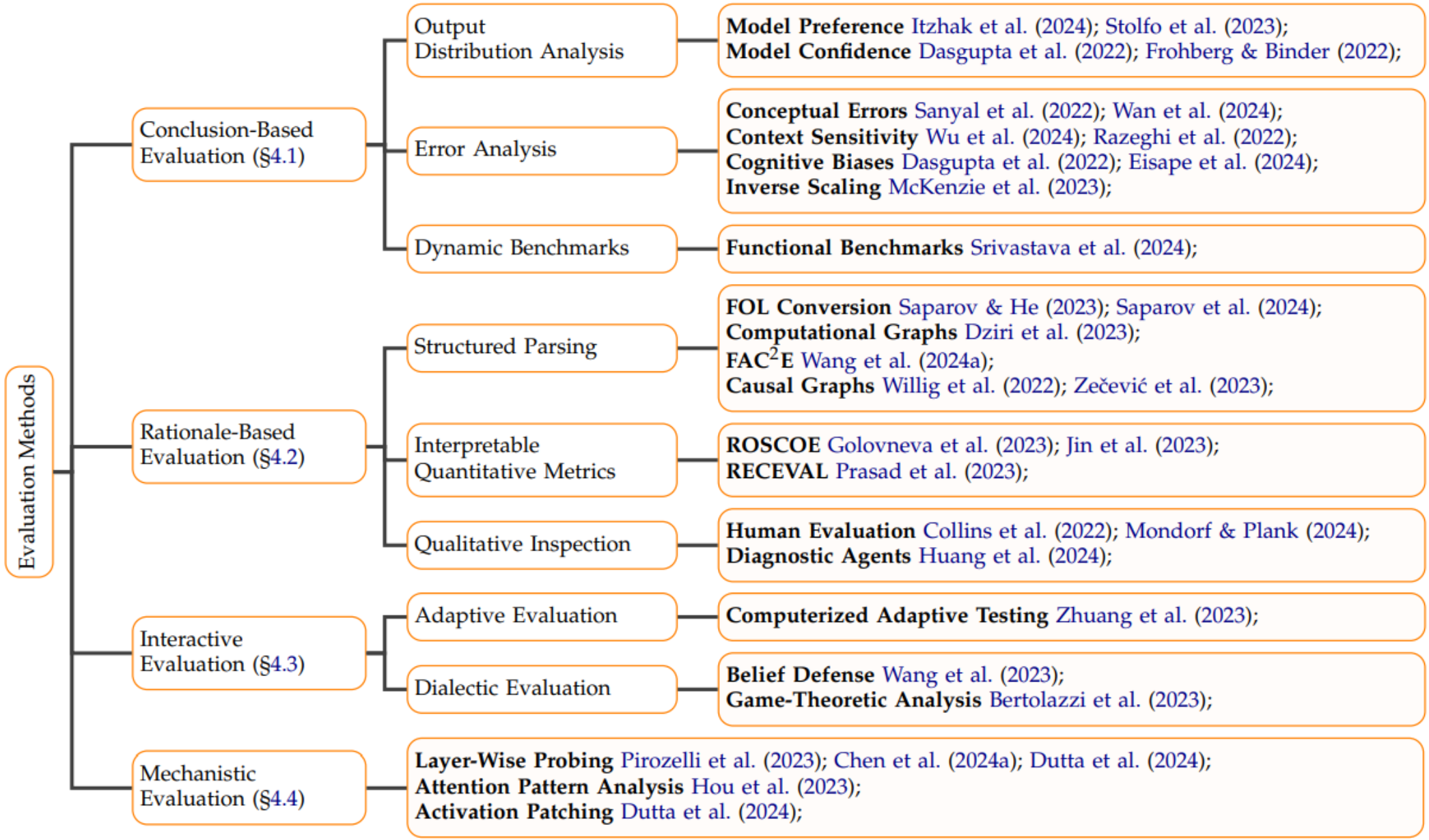
分布外の設定：
- 前提となる知識の入力順序を入れ替えたり
- マルチステップの推論で途中で間違えると回復できなかったり
- 無関係な事実が混入したにハルシネーションを起こしたり
- 非単調性に対応できない≒関連する新しい事実が入ってきたときに予測を変えられなかったり
  - {crow, peacock, **rabbit**} → bird （rabbitが加わると尤度が下がってほしいがLLMは下がらない）

品川の感想：
- 結局訓練が足りてないだけとも言える？あらゆるパターンを概ね訓練で網羅できたら解決する話？
- 全パターンを数えられる場合に、何%までカバーされると十分かという話に興味がある

後半の話
RQ2: LLMにおけるReasoning Behaviorを評価する方法のトレンドは何か？

- 評価方法を細目ごとに分類
- 一長一短

Reasoning Behaviorの評価方法一覧



Conclusion-Based Evaluation

結論のみを評価、推論の過程は重視しない

エラー分析	Sanyal et al. (2022) ：モデルが論理的な概念理解において誤りを犯すケースを確認 Dasgupta et al. (2022) ：認知バイアスの影響が大きい Wu et al. (2024) ：タスク文脈が変わるとモデルの回答が大きく影響される
出力分布の分析	Itzhak et al. (2024) ：指示付きチューニングされたモデルが新たなバイアスを示す Frohberg & Binder (2022) ：モデルが特定の結論に対して高い確信度を示す傾向
dynamic benchmarks	データの中身を動的に変化させて評価させることで、汎化能力を見る Srivastava et al. (2024) ：問題の構成を変化させて、モデルが単なる問題の記憶に依存していないかどうかを評価

Conclusion-Based Evaluationの課題

推論過程を無視していることによる限界

そもそも推論過程と結論（回答）がマッチしてない場合がある
（スコアが上がっても、中身を見てみると・・・🙈）

既存のベンチマークがLLMsのトレーニングデータに含まれている場合、正答率自体は本来より上振れしてしまう

関連：LLMが訓練データの回答をそのまま出してしまう問題
Balloccu et al. (2024) 、Xu et al. (2024)

モデルが回答に至るまでの推論の過程（根拠や説明）に注目し、論理の一貫性や妥当性を評価

Structure parsing	Saparov & He (2023) ：モデルの推論過程を一階述語論理に変換して論理的な妥当性を評価 Dziri et al. (2023) ：モデルの推論過程を計算グラフにパースし、各ステップの合理性を分析
Interpretable quantitative metrics	reasoning taskにおける根拠の意味的なアラインメントを評価 ROSCOE (Golovneva et al., 2023) RECEVAL (Prasad et al., 2023)
人手評価 or診断エージェント	構造化されてない評価は人手で定性的な評価をする Mondorf & Plank (2024) ：人間の評価者がモデルの推論を観察して評価

Rationale-Based Evaluationの課題

大部分は人手評価に頼る面が大きい
専門的な知識が要求する場合もあり、スケールさせることが難しい

Interactive Evaluation

モデルと対話を行いながら評価を進める
モデルが特定の推論過程に対してどのように応答を変えるかを評価できるのでより詳細な分析ができる

Adaptive Evaluation	Zhuang et al. (2023)：モデルの応答に応じて質問を動的に選ぶ
Dialectic Evaluation（弁証的評価）	LLMの結論に対して反論や質問を投げかけることで、LLMが自己の推論をどのように守るか、あるいは修正するかを観察 Wang et al. (2023) ：モデルに対して反論する形で対話を行い、モデルが自己防衛のためにどのような推論や説明を行うかを評価
ゲーム理論的分析	ゲーム理論に基づくシナリオを使って、モデルが競争的または協力的な場面でどのように推論を行うかを評価 Bertolazzi et al. (2023) ：20の質問ゲーム形式を用い、モデルが情報を探り当てる戦略を観察

Interactive Evaluationの課題

コストが高く、スケールを拡大して自動化するのが難しい
標準化が困難で、対話の設定や評価基準のばらつきによって再現性が低くなる
👆めっちゃわかる

入出力だけでなく、モデル内部の状態を分析して推論のメカニズムに注目する

Layer-Wise Probing	層ごとに知識の組み合わせの二値分類など Pirozelli et al. (2023) ：モデルの上位層がより複雑な推論や知識処理に関与する
attentionの分析	モデルの推論時にどのような単語やフレーズに注意が向けられているかを分析 Hou et al. (2023) ：GPT-2やLLaMAモデルでは、推論の過程で注意が段階的に移行、情報処理が階層的に進行していることがわかる
Activation Patching	特定の層やユニットのactivationを操作することで、モデル内部の推論過程を分析 Dutta et al. (2024) ：CoTプロンプト入力に対してLLMの中間層のactivationを操作するとLLMの応答が変化する

Mechanistic Evaluationの課題

計算コストが高く、結果の解釈が難しいので強い主張もむずかしい

- 計算コストが高い（probingのために層ごとに分類器を学習したり？）
- モデルの特定のタスクのみに適用できる場合が多く汎用性に欠ける
- 得られる結果の解釈が難しい

後半のまとめ

Conclusion-based：定量評価でスケールできるが、推論過程を無視
Rationale-base, Interactive：人手評価を実施するため高コスト（スケールしない、実験コントロールが大変）
Mechanistic：特定のタスク・モデルで評価する傾向があり、結果の解釈も難しいので一般化した主張が難しい

Evaluation Method	Advantages	Disadvantages
Conclusion-based evaluation	Allows for controlled setups Provides metrics for comparison Easy to automate and scale Easy to reproduce	Limited insights Less reliable
Rationale-based evaluation	Offers more nuanced insights More robust in certain scenarios	Difficult to automate and scale Might require expert interpretation
Interactive evaluation	Highly flexible Customizable to model behavior	Expensive Difficult to automate and scale Less standardized and reproducible
Mechanistic evaluation	Identifies features or circuits responsible for specific behaviors Supports direct interventions on model internals	Findings may not generalize across tasks or models Results may be hard to interpret Compute-intensive

全体の総括

様々なReasoningタスクにおけるLLMの挙動（Reasoning Behavior）に注目した

- Logical Reasoning (Deductive Reasoning, Inductive Reasoning, Abductive Reasoning)
- Mathematical Reasoning
- Causal Reasoning

論文のメインの主張：LLMは結局訓練データに類似するパターンには強いが、分布外の設定には弱い（確率的オウムの域を出ていない）

評価手法の傾向

- 定量評価：
  - LLMの入力や内部状態をいじって出力・出力分布の変化を見る
  - Deductive Reasoningでは根拠に基づいた定量評価が行われている（structure parsing, interpretable quantitative metrics）
  - probingで層ごとに知識の組み合わせに対する二値分類を適用する
- あとは定性評価を人手で頑張る！

感想：
オープンなベンチマークなど飾り、地道に多角的に分析するのが大事感が伝わってきた・・・