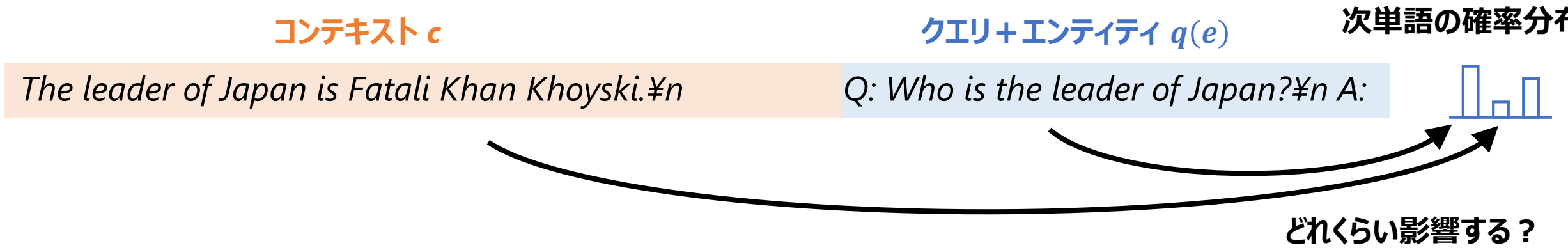
Context versus Prior Knowledge in Language Models

Kevin Du[ð] **Vésteinn Snæbjarnarson**[ᚴ] **Niklas Stoehr**[ð]
Jennifer C. White[ŋ] **Aaron Schein**[ɔ] **Ryan Cotterell**[ð]
[ð]ETH Zürich [ᚴ]University of Copenhagen
[ŋ]University of Cambridge [ɔ]The University of Chicago

紹介者：豊田工業大学 知能数理研究室　辻村 有輝

この研究は？：LLMのコンテキスト・クエリ（質問）がどの程度モデルの出力をコントロールするかを評価するスコアを提案

- 特にコンテキスト・エンティティごとの影響力の違いに注目して評価



背景

- 事前学習時に獲得したprior knowledgeと，推論時に与えられるコンテキストは矛盾しうる
  - ここで扱う"知識"は，エンティティと，エンティティ同士の関係とする
  - 本研究（や先行研究）ではentity biasとも呼んでいる
- 矛盾によるモデル出力への影響の理解は幻覚の低減において重要

- 矛盾発生時の影響調査に関する先行研究
  - (Longpre et al., 2021)：QAタスクにおいて，学習時に使ったデータ中のエンティティを別のエンティティに置き換えて，どの程度元のままの出力を答えてしまうかを調査
  - (Pezeshkpour 2023)：モデルへ知識を追加する際の，追加前後での出力変化をKLダイバージェンスで評価（本研究とかなり似ている）．
    - 知識追加の方法はprompt or fine-tuning

モチベーション

著者らの仮説：先行研究は調査対象のコンテキストやエンティティのそれぞれを同等に扱っているが，実際は**コンテキスト・エンティティの種類ごとに影響力が異なるのでは？**

- 強い事前知識があるエンティティなら，他のエンティティより影響力も強いはず
  - ***Harry*** *hugged* ***Voldemort****. How friendly are* ***Harry Potter*** *and* ***Lord Voldemort****?*
  - ***Susie*** *hugged* ***Alia****. How friendly are* ***Susie*** *and* ***Alia****?*
    - ハリーポッターとヴォルデモートの関係性によって出力により大きな影響を与えるはず
- コンテキストの形式によっても影響力が変わるのでは？
  - *The capital of Japan is Osaka. Where is the capital of Japan?*
  - ***Definitely****, the capital of Japan is Osaka. Where is the capital of Japan?*

➡ 本研究ではコンテキスト・エンティティの違いに注目した評価スコアを考え，実際に評価する

コンテキスト，クエリ＋エンティティの例 1/3

この研究での実験の流れ

① YAGOから抽出した122種の関係を用意

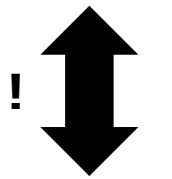
② 関係ごとにそれぞれ100個のエンティティと正解ペアを用意

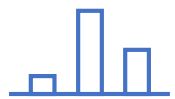
③ 関係に関するコンテキストとクエリを用意

④ コンテキストとクエリをLLMに与えて，その次のトークンの確率分布を調査

The leader of Japan is Fatali Khan Khoyski.¥n	*Q: Who is the leader of Japan?¥n A:*
コンテキスト c	クエリ＋エンティティ $q(e)$
Definitely, the leader of Japan is Fatali Khan Khoyski.¥n	*Q: Who is the leader of Japan?¥n A:*

比較！





コンテキスト，クエリ＋エンティティの例 2/3

① YAGOから抽出した122種の関係を用意

- *leader_of*, *made_of*, *member_of*, …

② 関係ごとにそれぞれ100個のエンティティと正解ペアを用意

- 50個は本物のエンティティ，残り50個はGPTによって生成させたfakeなエンティティ
- *leader_of*の例： (*Japan*, *Fumio Kishida*), (*Tokyo*, *Yuriko Koike*), …

コンテキスト，クエリ＋エンティティの例 3/3

③ 関係specificなコンテキストとクエリを用意

- **コンテキストはbase, assertive, negationの3種**
  - 関係=*leader_of* の時のコンテキストのテンプレート：
    - base : *The learder of {entity1} is {answer1}.*
    - assertive: *Definitely, the learder of {entity1} is {answer1}.*
    - negation: *The leader of {entity1} is not {answer1}.*
  - *{entity1}* と *{answer1}* は独立にサンプル（*The learder of Japan is Biden.*がありうる）
- **クエリはopen question 2種, closed question 2種**
  - *Q: Who is the leader of {entity2}?¥nA:*
  - *The leader of {entity2} is*

  } open

  - *Q: Is {entity2} the leader of {answer2}?¥nA:*
  - *Q: Is {answer2} led by {entity2}?¥nA:*

  } closed

  - *{entity1}*と*{entity2}*は独立
    - 実験設定上*{entity1}*=*{entity2}*になるケースが必ず発生
  - *{entity2}*と*{answer2}*は正しいペア

④ コンテキストとクエリをLLMに与えて，その次のトークンの確率分布を調査

コンテキスト，クエリ＋エンティティの例 1/3 (再掲)

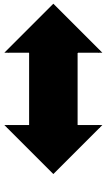
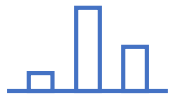
この研究での実験の流れ

① YAGOから抽出した122種の関係を用意←leader_ofなど

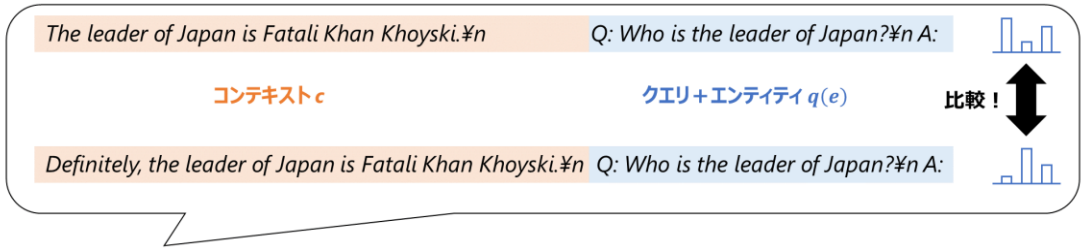
② 関係ごとにそれぞれ100個のエンティティと正解ペアを用意←(Japan, Fumio Kishida) など

③ 関係に関するコンテキストとクエリを用意←コンテキスト3種，クエリ2x2種

④ コンテキストとクエリをLLMに与えて，その次のトークンの確率分布を調査

コンテキスト c	クエリ＋エンティティ $q(e)$	
The leader of Japan is Fatali Khan Khoyski.¥n	*Q: Who is the leader of Japan?¥n A:*	
		比較！ 
Definitely, the leader of Japan is Fatali Khan Khoyski.¥n	*Q: Who is the leader of Japan?¥n A:*	

提案手法



相互情報量ベースの評価スコアを2種類提案

- スコアはコンテキスト，クエリ＋エンティティを入れ替えながらモデル出力を観測することで得られる

- Persuasion Score：$\psi(c, q(e))$
  - **コンテキスト c のモデル出力への影響力**を示すスコア
- Susceptibility Score：$\chi(q(e))$
  - **クエリ＋エンティティ $q(e)$ のモデル出力への影響力**を示すスコア

これらのスコアを用いてコンテキストやエンティティのバイアスなどがどの程度モデルに内在するか評価していく

提案手法｜Persuasion Score

Persuasion Score $\psi\big(c, q(e)\big)$：**コンテキスト c のモデル出力への影響力**を示すスコア

$$\begin{aligned}\psi\big(c, q(e)\big) &\triangleq \mathrm{I}(C = c; A | q(E) = q(e)) \\ &= \sum_{a \in \Sigma^*} p(a|c, q(e)) \log \frac{p\big(a|c, q(e)\big)}{p\big(a|q(e)\big)} \\ &= \mathrm{KL}\big(p(A|c, q(e)) || p(A|q(e))\big)\end{aligned}$$

- 注意点
  - クエリ＋エンティティ $q(e)$ にも依存
- スコアが表す意味
  - **大きいほどそのコンテキストがモデルの出力をコントロールできる（説得力がある）**
  - $\psi = 0$ (lower bound) ⇒ コンテキストは一切出力に影響を与えない

提案手法｜Susceptibility Score

Susceptibility Score $\chi(q(e))$: クエリ+エンティティ$q(e)$のモデル出力への影響力を示すスコア

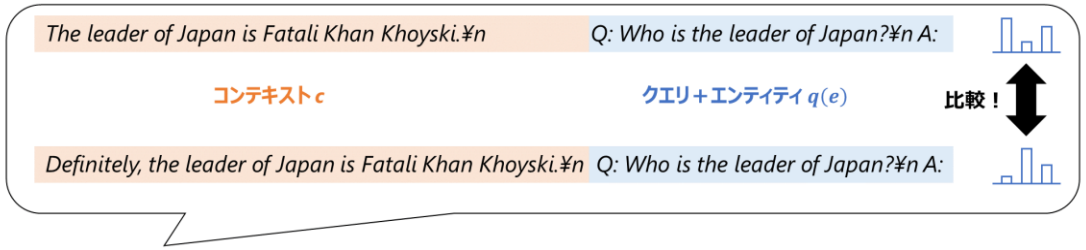
$$\begin{aligned}\chi(q(e)) &\triangleq \mathrm{I}(C; A|q(E) = q(e)) \\ &= \textstyle\sum_{c\in\Sigma^*} p(c)\psi(c, q(e)) \\ &= \mathrm{H}(A|q(E) = q(e)) - \mathrm{H}(A|C, q(E) = q(e))\end{aligned}$$

$\psi(c, q(e))$からコンテキストの条件付けが無くなっている

$$\psi(c, q(e)) \triangleq \mathrm{I}(C = c; A|q(E) = q(e))$$

- スコアが表す意味
  - **小さいほど**クエリ＋エンティティ $q(e)$ によってモデルの出力が決まり（バイアスが強い）
    大きいほどその $q(e)$ に対する出力はコンテキストからの影響を受けやすい
  - $\chi = 0$ (lower bound) ⇒コンテキストは出力に影響を与えず $q(e)$ のみで出力が決定
  - $\chi = \mathrm{H}(A)$ (upper bound) ⇒ $q(e)$ は出力に影響を与えずコンテキストによってのみ決定

提案手法（再掲）



相互情報量ベースの評価スコアを2種類提案

- スコアはコンテキスト，クエリ＋エンティティを入れ替えながらモデル出力を観測することで得られる

- Persuasion Score：$\psi(c, q(e))$
  - **コンテキスト c のモデル出力への影響力**を示すスコア
- Susceptibility Score：$\chi(q(e))$
  - **クエリ＋エンティティ $q(e)$ のモデル出力への影響力**を示すスコア

これらのスコアを用いてコンテキストやエンティティのバイアスなどがどの程度モデルに内在するか評価していく

実験

- 実験内容
  - 提案するPersuasive scoreとSusceptible scoreの妥当性の確認
  - Persuasiveなコンテキストの傾向
  - Susceptibleなエンティティの傾向の調査
- 実験設定
  - YAGO上の122種の関係について各600コンテキストと400クエリを用意
    - 関係ごとにreal 50種, GPT生成のfake 50種の計100エンティティを用意
    - 各エンティティごとに6コンテキスト（とanswer）を用意 ➡ 計600コンテキスト
      - コンテキストはbase, assertive, negationの3種類でevenly distributed
    - さらに各エンティティごとにopen, closed question各2種のクエリを用意 ➡ 4x100クエリ

    ➡ コンテキストにはほとんど (594/600) のケースでクエリと異なるエンティティが出現することになるが, 6コンテキストではクエリと同一のエンティティが用いられることになる
- モデル：Pythia (deduped Pileで学習されたLLM)
  - サイズ：70m, 410m, 1.4b, 2.8b, 6.9b, 12b (8-bit quantized)

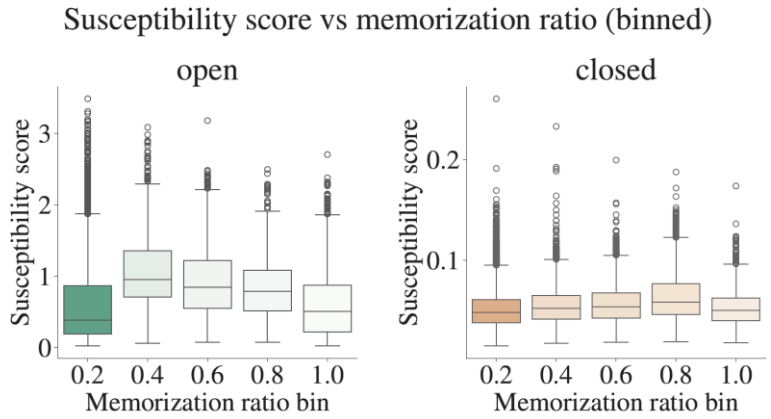
実験｜提案スコアの妥当性の確認

Persuasion score

- コンテキストを入れ替えながら生成を行い，出力とPersuasion scoreを記録
- コンテキスト中のエンティティを答えたケースは，本来の正解エンティティを答えるケースより高いPersuasion scoreを得ると期待して有意差検定を実施.
  - Open question：モデル出力を変動させるコンテキストのうち，**59%が有意に高い**結果
  - Closed question：**34%が有意に高い**

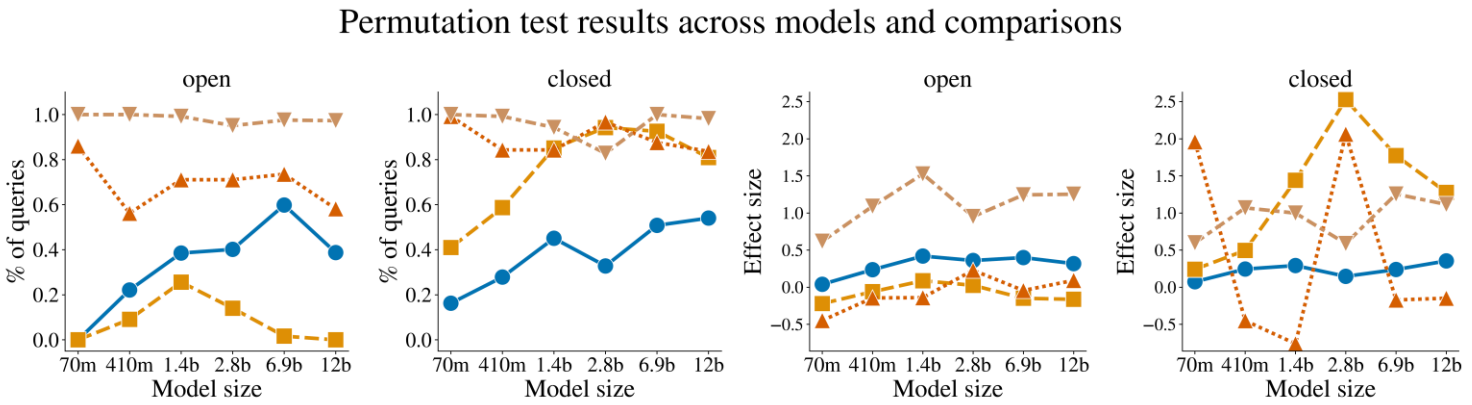
Susceptibility score

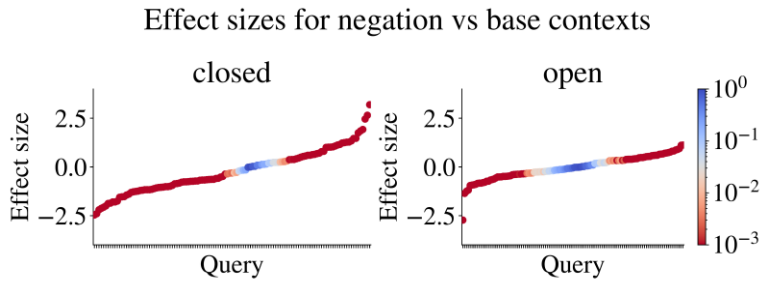
- Susceptibility scoreとmemorization ratio (MR)を比較
  - MR：コンテキスト変動時にどの程度の割合で元々の正解エンティティのまま回答するか
- 期待：MRが高い↔Susceptibility scoreは低い（いずれもコンテキストに影響されにくいことを表す）
- 結果：**upper bound**が抑えられる



実験｜Persuasiveなコンテキストの傾向

- **Relevant**：コンテキスト中に出現するエンティティとクエリ中のエンティティが同じ
  - 結果：ほとんどのクエリでエンティティが異なるケースより有意に高い
- **Assertive**：*"Definitely, the leader of Japan is Biden."* のようなコンテキスト
  - 結果：closedではbaseコンテキストより有意に高いクエリが多いが，openでは有意差がほとんど出ない
- **Negation**：*"The leader of Japan is not Biden."* のようなコンテキスト
  - **両側検定**でbaseに対し多くに有意差が出るが，**効果量に一貫性がない**



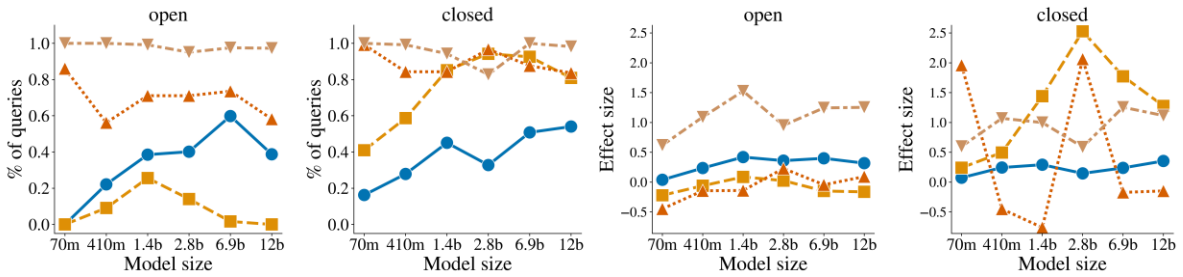


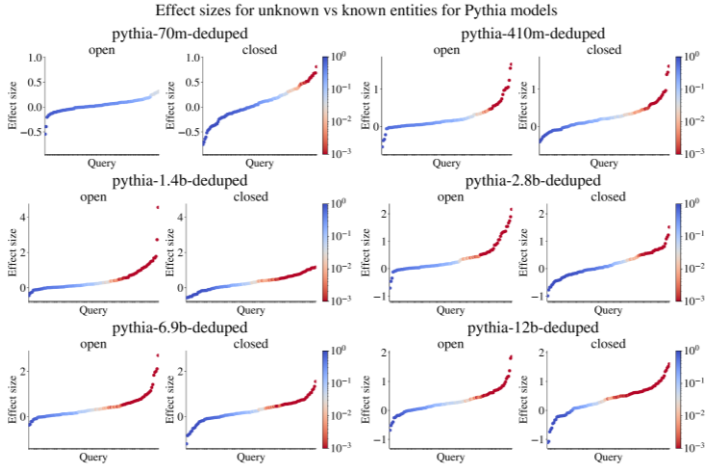
6.9bモデルでのnegation vs baseコンテキストのクエリごとの有意差検定・効果量．色はp値を表す

実験｜Susceptibleなエンティティの傾向 1/3

- **検証したい仮説：pretraining中に遭遇したことのあるエンティティはSusceptibility scoreが低くなる（コンテキストに影響されにくくなる）のではないか**
- 実験：realエンティティとGPT生成のfakeエンティティのSusceptibility scoreを比較
- 結果：モデルサイズが大きくなるにしたがって実際のエンティティの方が大きな効果量になる傾向が増し，有意差の認められるクエリの割合も増える

図中青の線●がreal/fakeの比較





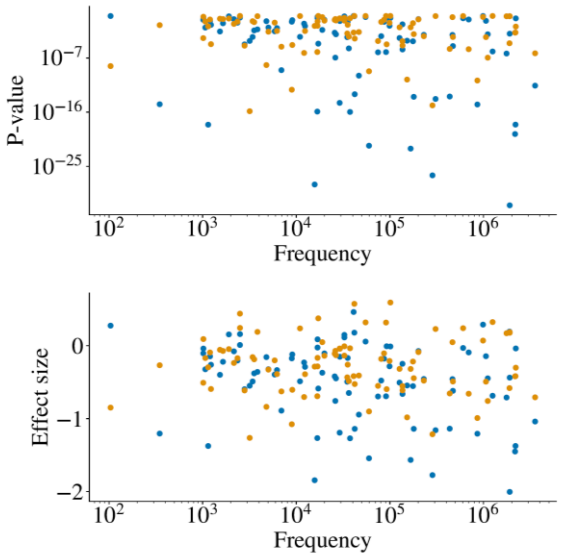
- しかしながら有意差が認められるのは多くて半数程度…
- モデルサイズが上がるとeffect sizeは平均的に上がるというよりは極端なものが増える？

実験｜Susceptibleなエンティティの傾向 2/3

- 先ページの結果：realエンティティとGPT生成のfakeエンティティのSusceptibility scoreの比較では多くとも半数程度しか有意差が認められなかった
- **結果に対する仮説：**
  - **出現回数が少ないとバイアスは発生しないのではないか**
  - **有意差が出なかったクエリはサンプリングしたrealエンティティの出現回数がたまたま少ないものばかりだったのではないか**

➡ Realエンティティの訓練データ（dedup Pile) 上での出現回数とSusceptibility scoreの相関関係を調査

- スピアマンの順位相関係数で評価
- **結果：open questionでは有意差のある負の相関 (ρ=-0.23) が得られたが絶対値としては小さく，closed questionでは有意差も認められなかった**



頻度ごとのreal/fakeエンティティ間のp値/効果量
青：open question　橙：closed question

実験｜Susceptibleなエンティティの傾向 3/3

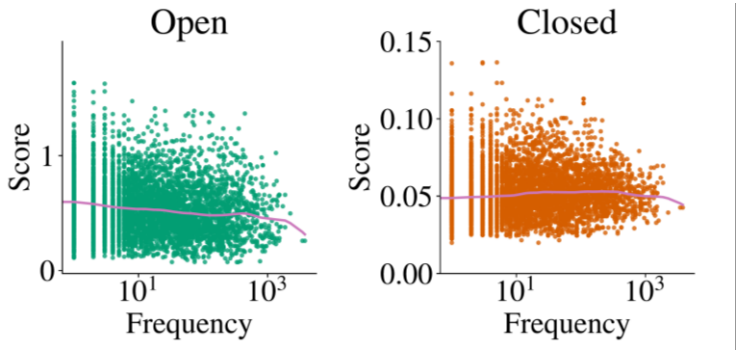
実験内容：単純なエンティティのみの出現回数ではなく，クエリ対象の関係にフォーカスしてSusceptibility scoreを調査

1) 訓練データ中のエンティティと関係の共起回数（文字列一致，50 word window）

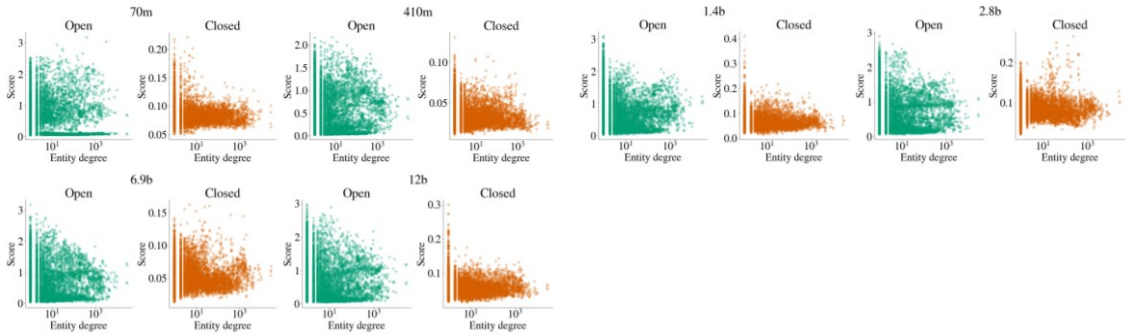
2) YAGO知識グラフ上でのクエリ対象の関係についての次数

- 期待：共起回数よりpreciseなはず

- 共起回数での評価：負の相関 (ρ=-0.23) を確認.
  - “upper boundが抑えられる”とのこと
  - closed側だけだと相関が出なそう・・・



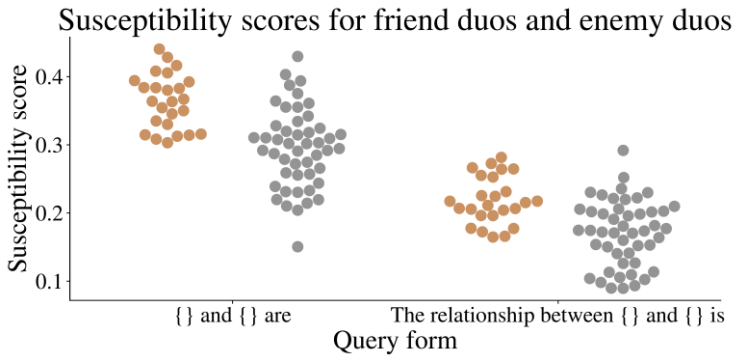
- 次数での評価：
  - こちらもupper boundが抑えられるとの分析だが・・・



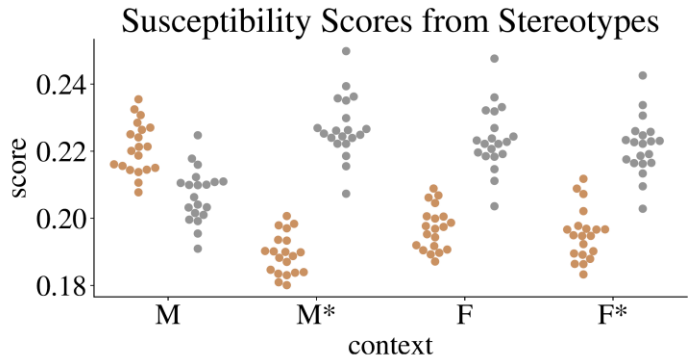
アプリケーション例

- Social Sciences Measurement
  - **LLMでデータ生成する際にどの程度entity biasesが影響するかを理解したい**

  例) *The relationship between {entity1} and {entity2} is* というクエリについて
  entity1 と *entity2* が友好関係か敵対関係かでどの程度モデルの出力が変わりうるかを Susceptibility scoreの計測によって認識できる (左図)
- Exploring Gender Bias
  - Susceptibility scoreからエンティティのgender biasを評価できる
  - ステレオタイプなコンテキスト＋男性or女性的エンティティのクエリでスコア計測 (右図)



灰色：Enemy duos　ゴールド：Friend duos



灰色：q(男性的エンティティ)　ゴールド：q(女性的エンティティ)

まとめ

相互情報量ベースの評価指標を2種類提案

- Persuasion Score：$\psi(c, q(e))$
  - コンテキスト c のモデル出力への影響力を示すスコア
- Susceptibility Score：$\chi(q(e))$
  - クエリ＋エンティティ $q(e)$ のモデル出力への影響力を示すスコア

結果

- Relevant, assertive, negation contextのPersuasion scoreへの影響を確認
- 学習時の出現回数，知識グラフ上での次数のSusceptibility scoreへの影響を確認
- しかしいずれも結果はあまりはっきりしない・・・

ほか所感

- 本文Limitationにも書いてあったが，提案手法はコンテキスト・クエリ直後のstepにおけるvocab全体の確率分布の変動から計算しているのはかなり大きな制限に感じる
  - 回答に直接関係するtokenのみへの影響や，その後のトークン列に渡る影響が不透明
  - もっと深く見ればもっと有益な（はっきりした）知見が得られるのでは？
    - ただしそうでもないのかもしれない；公式レポジトリには特定トークンだけでスコアを計測する機能が実装されているが，対応する実験結果が論文中で報告されていない

- Pythia以外のモデルの評価結果がないのはモデルの評価手法提案としてどうなのか？
  - 計算量が大きくて大変とは書いていたが・・・
  - Instruction tunedなモデルだとどうなる？