

# Context versus Prior Knowledge in Language Models

Kevin Du<sup>δ</sup> Vésteinn Snæbjarnarson<sup>f</sup> Niklas Stoehr<sup>δ</sup>

Jennifer C. White<sup>‡</sup> Aaron Schein<sup>‡</sup> Ryan Cotterell<sup>δ</sup>


<sup>δ</sup>ETH Zürich <sup>f</sup>University of Copenhagen

<sup>‡</sup>University of Cambridge <sup>‡</sup>The University of Chicago

kevin.du@inf.ethz.ch vesn@di.ku.dk niklas.stoehr@inf.ethz.ch  
jw2088@cam.ac.uk schein@uchicago.edu ryan.cotterell@inf.ethz.ch

## Abstract

To answer a question, language models often need to integrate prior knowledge learned during pretraining and new information presented in context. We hypothesize that models perform this integration in a predictable way across different questions and contexts: models will rely more on prior knowledge for questions about entities (e.g., persons, places, etc.) that they are more familiar with due to higher exposure in the training corpus, and be more easily persuaded by some contexts than others. To formalize this problem, we propose two mutual information-based metrics to measure a model’s dependency on a context and on its prior about an entity: first, the *persuasion score* of a given context represents how much a model depends on the context in its decision, and second, the *susceptibility score* of a given entity represents how much the model can be swayed away from its original answer distribution about an entity. We empirically test our metrics for their validity and reliability. Finally, we explore and find a relationship between the scores and the model’s expected familiarity with an entity, and provide two use cases to illustrate their benefits.

 <https://github.com/kdu4108/measureLM>

## 1 Introduction

Language models have displayed remarkable abilities to answer factual queries about entities, suggesting that they encode knowledge about these entities learned during pretraining (Petroni et al., 2019; Brown et al., 2020; Roberts et al., 2020; Geva et al., 2021). For prompts that extend a question with additional information or context, the model can draw on both its prior knowledge and the additional context to answer the query (Kwiatkowski et al., 2019; Joshi et al., 2017; Berant et al., 2013; Kasai et al., 2023). While previous research has investigated how often a model will rely on prior knowledge over conflicting contextual information in answering questions (Longpre et al., 2021), we

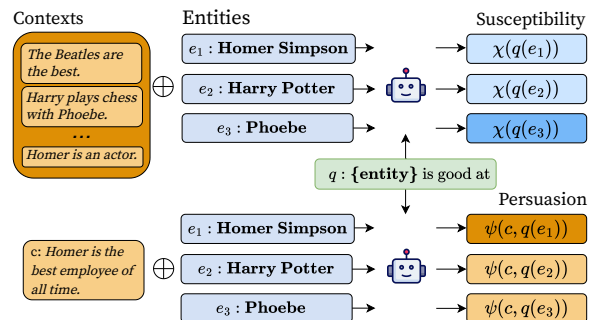


Figure 1: In answering a given query, a model may be more *susceptible* to context for some *entities* than others, while some *contexts* may be more *persuasive* than others (as indicated in this figure by color darkness in the rightmost column). We introduce mutual information-based metrics to evaluate how much impact the context has relative to the prior knowledge of a model.

hypothesize that models will not behave identically for all contexts and entities. For example, if a language model is prompted with *Harry hugged Voldemort. How friendly are Harry Potter and Lord Voldemort?*, we might expect the prior knowledge learned from training data describing the rivalry between these two characters to significantly influence the model’s answer. However, if the model lacks a strong prior on, say, *Susie* and *Alia*, then we might expect its answer to be primarily context-driven when prompted with *Susie hugged Alia. How friendly are Susie and Alia?*

We formalize this problem through the lens of evaluating the change in a model’s answer distribution for different contexts and entities. We present two mutual information-based metrics that allow us to explore differences in the effect of specific contexts on model behavior for different entities. The **persuasion score** of a given context measures how much a model’s answer distribution is affected by the context when prompted with a particular query about a given entity. The **susceptibility score** of a given entity measures how much the model’s answer distribution can be swayed for a particular query about that entity, marginalized over all contexts. Given their basis in mutual information, our metrics are natural operationalizations of per-

suasion and susceptibility. Furthermore, we offer empirical evidence of the validity and reliability of these measures by comparing them against similar metrics and showing their robustness to paraphrases and different samples.

To study how language models behave for different contexts and entities, we create a synthetic dataset of queries covering 122 topic areas extracted from the YAGO knowledge graph (Suchanek et al., 2007), entities extracted from YAGO and generated with GPT-4 (OpenAI, 2023), and contexts constructed with different qualities, e.g., relevancy and assertiveness. We apply our new metrics to Pythia models ranging from 70m to 12b parameters (Biderman et al., 2023) to find evidence that relevant contexts are consistently more persuasive than irrelevant ones, and assertive contexts are more persuasive than less assertive ones for yes–no questions. In a deep dive into one model, we find evidence that entities with high expected familiarity, as measured by both training data frequency and entity degree statistics in the YAGO knowledge graph, have lower susceptibility scores.

We further conduct case studies to show how these metrics could be useful in applied settings. In a study on friend–enemy stance measurement, we find evidence that enemy duos are less susceptible than friend duos. Applying our metrics to gender bias, we find evidence for a difference in susceptibility between stereotypically masculine and feminine names for gender-biased contexts. Through this, we show how our proposed metrics can be used to better analyze the effects of context and prior knowledge, with the potential for application toward greater control over model behavior.

## 2 Context and Prior Knowledge

Much prior work has noted that language models develop biases about entities during training and investigated the tension between this entity bias and additional information provided in the context.

### 2.1 Entity Bias

Loosely defined across studies as model bias from spurious correlations between entity mentions and a target characteristic in the training data, **entity bias** has been mainly examined in relation extraction (RE), where a model extracts relationships between entities from text (Zelenko et al., 2002; Zhou et al., 2005). Different studies on RE have attempted to either mitigate (Zhang et al., 2017,

2018; Peng et al., 2020; Wang et al., 2022) or leverage (Yamada et al., 2020; Zhou and Chen, 2022) entity bias for improved performance on the task. Wang et al. (2022) note that entities can carry both useful semantic information about their roles, e.g., whether the entity is a person or a place, and spurious information which can bias the model toward relations not mentioned in the target sentence, e.g., the model may link *Switzerland* with the relation *countries\_of\_residence* at inference time for a sentence that does not mention this relation, due to their frequent co-occurrence in the training data. Additional studies on machine reading comprehension (MRC) tasks, e.g., SQuAD (Rajpurkar et al., 2016), TriviaQA (Joshi et al., 2017), and NaturalQuestions (Kwiatkowski et al., 2019), have found substituting different entity names can result in meaningful changes in model predictions and overall evaluation performance (Yan et al., 2022).

### 2.2 Context and Entity Bias

The existence of entity bias naturally raises the question of how it interacts with context to shape a model’s response. Several papers (Longpre et al., 2021; Chen et al., 2022; Xie et al., 2023) approach this by inducing and exploring **knowledge conflicts**, where a context preceding a query proposes information that conflicts with a model’s prior knowledge about that query. They measure the model’s reliance on pretrained entity bias and context by computing the **memorization ratio**: the proportion of knowledge conflict examples for which the model maintains its answer from prior knowledge. Pezeshkpour (2023) measures a model’s prior knowledge of a fact by comparing the entropy of a queried model’s answer distribution before and after stating the fact in context. Several studies further propose interventions to reduce entity bias and favor in-context information. These approaches include prompting (Zhou et al., 2023; Onoe et al., 2023), modifying training data (Wang et al., 2023), fine-tuning (Li et al., 2023), and neuron-level interventions (Yu et al., 2023) at inference time.

However, the metrics used to quantify model reliance on context and entity bias in these papers—excepting Pezeshkpour (2023)—are limited in several ways. First, most previous work does not develop *entity-specific* or *context-specific* metrics for the strength of the entity bias or context persuasiveness. Instead, their metrics produce only

a single number to summarize the model’s overall reliance on entity bias. Second, their setups are limited to adversarial cases in which a context is chosen to go against the entity bias. Indeed, we may well wish to measure the interplay of context and entity bias in other cases, such as when context reinforces entity bias or when they do not clearly disagree. Therefore, we seek to define a metric that measures how much the model depends on a given context or entity with rigorous, theoretically grounded interpretations.

### 3 Our Formalization

We now formalize the problem setting in which we define metrics for the susceptibility of an entity and the persuasiveness of a context for a given model. Let  $\Sigma$  be an alphabet. Consider a language model  $p_M$  over  $\Sigma$ , i.e.,  $p_M$  is a distribution over the Kleene closure  $\Sigma^*$ . Furthermore, we assume that  $p_M$  was estimated from a corpus  $\mathcal{D} \subset \Sigma^*$ . Let  $\mathcal{E} \subset \Sigma^*$  be the subset of strings that could correspond to string representations of entities.<sup>1</sup> Let  $Q = \{q_n\}_{n=1}^N$  be a set of  $N$  **slotted query templates** of type  $q_n: \mathcal{E} \rightarrow \Sigma^*$  that fill the slot with the argument to each  $q_n$ . We can think of a query template  $q \in Q$  as slotting an entity into a query, e.g., slotting the entity *Slovenia* into the query template  $q(e) = \textit{The capital of } e \textit{ is}$  produces the string *The capital of Slovenia is*.

Now, consider three random variables. First, let  $C$  be a  $\Sigma^*$ -valued random variable that stands for a context. Second, let  $A$  be a  $\Sigma^*$ -valued random variable whose values are answers. Let  $E$  be a  $\mathcal{E}$ -valued random variable. The pushforward  $q(E)$ , then, is a  $\Sigma^*$ -valued random variable over slottings of entities according to query  $q \in Q$ . The random variables  $C$ ,  $q(E)$  and  $A$  are jointly distributed according to the following probability distribution

$$p(C = c, q(E) = q(e), A = a) \hat{\propto} p_M(cq(e)a), \quad (1)$$

where  $cq(e)a \in \Sigma^*$  is string concatenation. To formalize persuasion and susceptibility, we make use of the joint distribution  $p$  heavily in the proceeding subsections.

#### 3.1 Persuasion Score

For each context  $c$  that is prepended to a query template with a given entity slotted in,  $q(e)$ , we wish to

<sup>1</sup>In practice, an entity can have different verbalizations, e.g., *Sherlock* and *Sherlock Holmes*. . . And, whether a specific string refers to an entity may very well depend on the context in which it occurs. Thus, our set  $\mathcal{E}$  is an approximation at best.

assign a **persuasion score**  $\psi$  to represent how successful that context is at altering a model’s answer distribution. This score depends on the specific queried entity  $e$ , because contexts themselves are often entity-dependent. Intuitively, a context’s persuasion score should measure how much the probability distribution of possible answers changes, averaged across all possible answers. More precisely, we define our persuasion score  $\psi(c, q(e))$  as the half-pointwise mutual information (half-PMI) between the context  $c$  and the answer random variable  $A$ , conditioned on the fixed query about an entity:

$$\begin{aligned} \psi(c, q(e)) &\triangleq I(C = c; A \mid q(E) = q(e)) \\ &= \sum_{a \in \Sigma^*} p(a \mid c, q(e)) \log \frac{p(a \mid c, q(e))}{p(a \mid q(e))} \\ &= \text{KL}(p(A \mid c, q(e)) \parallel p(A \mid q(e))), \end{aligned} \quad (2)$$

where  $p(A \mid c, q(e))$  and  $p(A \mid q(e))$  can be derived by marginalizing and conditioning Eq. (1).

The persuasion score of a context can then be interpreted as the degree (in nats) to which the context was able to change the model’s answer distribution when prepended to a query. When the persuasion score is at its lower bound of 0 nats, it indicates the context is completely *unpersuasive*, i.e., it did not change the model’s answer distribution at all. Contexts with higher persuasion scores change the answer distribution more, which is consistent when viewed through the lens of KL-divergence.<sup>2</sup>

#### 3.2 Susceptibility Score

For a given query template applied to an entity  $q(e)$ , we further wish to assign a susceptibility score  $\chi$  to  $q(e)$  which represents how easy it is to change a model’s answer distribution. Intuitively, the susceptibility score should measure how much a model’s answer distribution to a query changes when prompted with additional context, averaged across all possible contexts and answers. More precisely, we define the susceptibility score  $\chi(q(e))$  as the mutual information between the context and answer random variables, conditioned on a fixed query about an entity:

$$\begin{aligned} \chi(q(e)) &\triangleq \sum_{c \in \Sigma^*} p(c) \psi(c, q(e)) \\ &= I(C; A \mid q(E) = q(e)), \end{aligned} \quad (3)$$

<sup>2</sup>We include details on further equivalencies of half-PMI and other concepts in information theory in App. A.

where  $p(c)$  is the marginal distribution over contexts. Equivalent to the difference in entropy  $H(A | q(E) = q(e)) - H(A | C, q(E) = q(e))$ , the susceptibility score represents the reduction in answer distribution uncertainty (in nats) when a query is preceded by a context. A high susceptibility score means the model is highly influenced by context for the query about that entity, with its upper bound of  $H(A)$  indicating that context fully determines the answer. A low score indicates the model’s response is robust to context, with its lower bound of 0 indicating no influence of context on the answer distribution. We can use susceptibility to answer the question “How much does the model’s answer depend on its entity bias?” with an information-theoretically grounded, interpretable scale based on the model’s full behavior.

### 3.3 Entity-Independent Persuasion Score

Persuasion scores can be further marginalized over entities. Analogous to our definition of the susceptibility score, we define the **entity-independent persuasion score** of a context as how much the log probability distribution of possible answers changes, averaged across all possible entities and answers. We describe this further in App. B.

## 4 Experiments

We now provide empirical evidence to further validate our metrics, characterize model behavior using the susceptibility and persuasion scores, and investigate reasons behind differences in susceptibility scores for different entities.

### 4.1 Setup

For each of the 122 different relations from the YAGO knowledge graph, we collect 100 entities<sup>3</sup> and construct 600 random contexts from relation-specific context templates such that each entity is mentioned in 6 contexts. The 600 contexts are evenly distributed between *assertive*,<sup>4</sup> *base*,<sup>5</sup> and *negation*<sup>6</sup> context types. We construct four query forms of each relation: two *closed* questions, i.e., yes–no, and two *open* questions.

Using these samples, we compute persuasion scores for each context according to Eq. (2) and

<sup>3</sup>50 are real entities (e.g., *Adele*) sampled from YAGO, and 50 are fake entities (e.g., *Udo König*) of the same entity class (e.g., *Person*) generated with GPT-4 (OpenAI, 2023).

<sup>4</sup>E.g., *Definitely, the capital of {entity} is {answer}*.

<sup>5</sup>E.g., *The capital of {entity} is {answer}*.

<sup>6</sup>E.g., *The capital of {entity} is not {answer}*.

susceptibility scores for each entity according to Eq. (3) for each of the four query forms for six Pythia models of different sizes<sup>7</sup> trained on the deduplicated Pile (Wolf et al., 2020; Detmers et al., 2022; Biderman et al., 2023). Due to computational constraints, we approximate the model’s answer distribution with the next-token distribution over the model’s vocabulary.<sup>8</sup> The detailed setup can be found in App. C.

## 4.2 Empirically Validating Our Metrics

### 4.2.1 Estimating Scores

Because persuasion and susceptibility scores both involve a countably infinite sum, we opt for a stochastic approximation scheme. Specifically, we construct a Monte Carlo estimator. We sample from a narrower set of constructed contexts and approximate the answer distribution with the next token distribution over the model’s vocabulary, as described in §4.1. We take a sample size of 600 contexts. While the Monte Carlo approximation itself results in a consistent estimator, the additional approximations mean we do not have a guarantee on the quality of the approximation as a whole. In Fig. 4, we exhibit the variance of our estimator across three random seeds, i.e., sampled sets of context.

### 4.2.2 Validating Persuasion Scores

**Convergent Validity.** According to existing measurement modeling methods (Loevinger, 1957; Messick, 1987; Jackman, 2008; Hand, 2004; Quinn et al., 2010; Jacobs and Wallach, 2021), observing a relationship between a new metric and existing ones would serve as additional evidence that the metric is meaningful. To this end, we explore whether contexts with higher persuasion scores tend to more successfully convince the model to agree with the context. Using the Pythia-6.9b-deduped model, we generate an answer for each prepended context to a query–entity pair from §4.1 and use simple string matching to map the answer to whether it agrees with the context, the original answer, or neither. We then apply a permutation test ( $k = 10000, \alpha = 0.05$  with the Benjamini–Hochberg (BH) correction (Benjamini and Hochberg, 1995)) for whether contexts that elicited context-concordant answers have higher persuasion scores than those that did not alter the

<sup>7</sup>70m, 410m, 1.4b, 2.8b, 6.9b, 12b (8-bit quantized)

<sup>8</sup>While it would be more precise to estimate the answer distribution by repeatedly sampling many model outputs, this is very computationally expensive.

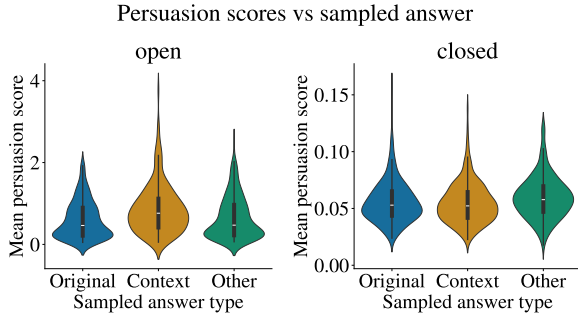


Figure 2: The  $x$ -axis represents bins for whether the model’s answer agreed with its prior, the context, or neither. For open (left) queries, the persuasion scores of contexts that persuaded the model to output an answer matching the context (*Context*) are higher than those of contexts that did not (*Original*, *Other*).

model from the original answer for each of the 122 query topics. Within a query topic, we run separate tests for open and closed queries, since the entropy of the answer distribution for closed queries tends to be much lower than the entropy for open queries. We find that for 59% of open queries, contexts that persuade the model to output the in-context answer have significantly higher persuasion scores than non-persuasive ones. Curiously, this behavior holds for only 34% of closed queries. We discuss this surprising result more in §6. We further show a summary of the mean persuasion scores for the different kinds of elicited answers in Fig. 2.

**Construct Reliability.** To show construct reliability, we provide evidence that persuasion scores do not strongly vary for inputs that differ only in phrasing, i.e., *query form*. For example, the persuasion score of *The capital of Slovenia is Oz.* should be similar when prepended to either *Is Slovenia’s capital Ljubljana?* or *Is Ljubljana the capital of Slovenia?*. We compute persuasion scores using the setup from §4.1 with two query forms for both closed and open queries, keeping contexts that appeared for the same query and entity for all seeds. We then compute the variance across the query forms to test for reliability. Fig. 4 shows strong evidence that the metric is reliable. The variance is very low across different closed query forms, as expected. While open query forms have higher variance, this is not unexpected because the question-answering query form, e.g., *Q: What is the capital of Slovenia?* is more specific than the sentence-completion form, e.g., *The capital of Slovenia is,* which has a broader set of plausible answers.

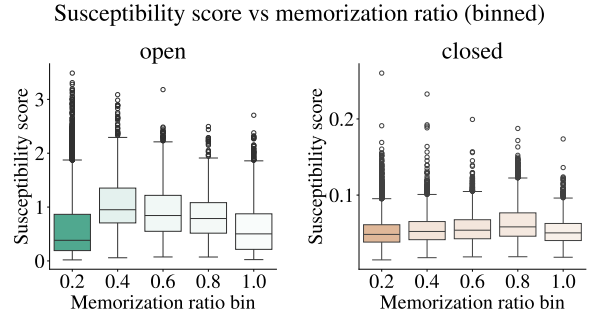


Figure 3: Susceptibility score ( $y$ -axis) against the MR divided into 5 bins between 0 and 1 ( $x$ -axis) for all entities and queries. The opacity represents the proportion of points in each bin. For both *open queries* (●) and *closed queries* (●), we see a decreasing upper bound between the MR and susceptibility score. While the quartiles of the open queries generally decrease (except for the lowest bin), the opposite occurs for closed queries.

### 4.2.3 Validating Susceptibility Scores

**Convergent Validity.** We compare susceptibility scores to per-entity memorization ratio (MR)<sup>9</sup> as further evidence for the meaningfulness of our metric, using the setup from §4.1 for the Pythia-6.9b-deduped model. Fig. 3 shows a decreasing upper-bounding relationship between susceptibility scores and MR for both open queries and closed queries. While not a straightforward correlation, this pattern supports the convergent validity of susceptibility scores and can be explained by the scores’ nature of measuring a difference in entropy. That is, if the model’s answer was often changed (low MR), this could correspond to a wide range of entropy change (low or high susceptibility), depending on the model’s confidence without the context. If the model’s answer was mostly unchanged (high MR), the entropy likely remained similar (low susceptibility). We discuss these results further in App. D.2.

**Construct Reliability.** Following the setup used in §4.2.2, we consider the same two query forms for both open and closed questions to test for reliability (low variance) for susceptibility scores. Fig. 4 shows strong evidence for the reliability of susceptibility scores. Similarly to the persuasion scores, the variance for both open and closed queries is very low across closed query forms; open query forms have higher variance because they have meaningfully different possible answers, as discussed in §4.2.2.

<sup>9</sup>Adapted from Longpre et al. (2021) to apply on a *per-entity* basis and describe it in more detail in App. D.

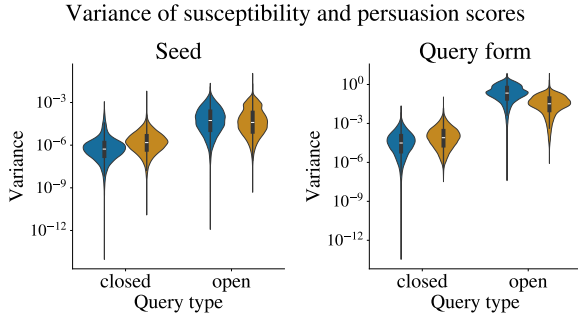


Figure 4: Summarizing across all 122 queries, we display the variance of *susceptibility scores* (blue) and *persuasion scores* (orange), across three random seeds (left) and across two query forms (right), and stratified for both closed and open queries ( $x$ -axis). The variance is very low across random seeds for both query types, and, for closed queries, across the specific query form. Variance is high for the different open query forms.

### 4.3 What Makes a Context Persuasive?

To better characterize model behavior with persuasion scores, we explore several tests for qualities that might distinguish more persuasive contexts from less persuasive ones: relevance, assertiveness, and negation.

#### 4.3.1 Relevance

**Experiment Setup.** We use the setup described in §4.1. For the relevance test, we consider a *relevant* context to be one that mentions the queried entity, and an *irrelevant* context as one that does not. We hypothesize that *relevant* contexts should be more persuasive than *irrelevant* ones. For each of the 122 queries, we find the mean persuasion score of the relevant contexts and irrelevant contexts and use a permutation test to determine whether the mean persuasion score for relevant contexts (across entities) is higher than the mean persuasion score for irrelevant contexts, using a significance level of  $\alpha = 0.05$  with the BH correction.

**Results.** As seen in Fig. 5 (▼), across most model sizes and relations, relevant contexts are significantly more persuasive than irrelevant contexts. Specifically, depending on the model, 95–100% of open queries and 83–100% of closed queries showed a significant result. We also see a trend in which, as model size increases, so too does the degree to which relevant contexts are more persuasive, as measured by the mean effect size. We summarize the significance test results and effect sizes for all models and queries in App. E.

#### 4.3.2 Assertiveness

**Experiment Setup.** Using a similar setup to §4.3.1, we explore whether *assertive* contexts are more persuasive than *base* ones by testing whether the mean persuasion score of the former group is greater than that of the latter for each query.

**Results.** Fig. 5 (■) shows that consistently across model sizes, assertive contexts tend to be more persuasive than base contexts for closed queries but not for open queries. Moreover, assertive contexts are most persuasive for medium-sized models such as 1.4b and 2.8b. We hypothesize that smaller models may be less persuaded by assertive contexts because they may be worse at integrating context into their answers. Larger models may be less persuaded because of their stronger prior knowledge/lower susceptibility to context, whether assertive or not. We highlight the significance test results and effect sizes for all queries for the Pythia-6.9b-deduped model in Fig. 6 and other models in App. E.

#### 4.3.3 Negation

**Experiment Setup.** We use the same setup as in §4.3.2 and explore whether *negation* contexts differ in persuasiveness from the *base* ones, using a two-tailed permutation test for each query.

**Results.** The permutation tests suggest evidence for a significant difference in 88% of the closed queries and 74% of the open queries for the Pythia-6.9b-deduped model. However, there is no consistent directional pattern; from Fig. 7, we see that negations are significantly more persuasive for some queries while significantly less persuasive for others. App. E shows a similar pattern for other models. Fig. 5 (▲) also shows that the smallest model is the most sensitive to being persuaded differently by negation vs base contexts. For closed queries, this may be due to potential spurious correlations or token biases between, i.e., seeing the word *not* and the model’s probability of outputting *No*.

#### 4.3.4 Comparing Context Qualities

For all models on open queries, the relevance of a context has the greatest effect on persuasion score compared to assertiveness or negation, as measured by effect size. This is consistent with our intuition that the model should be more sensitive to whether the queried entity is mentioned in the context than to other context features. However, we

## Permutation test results across models and comparisons

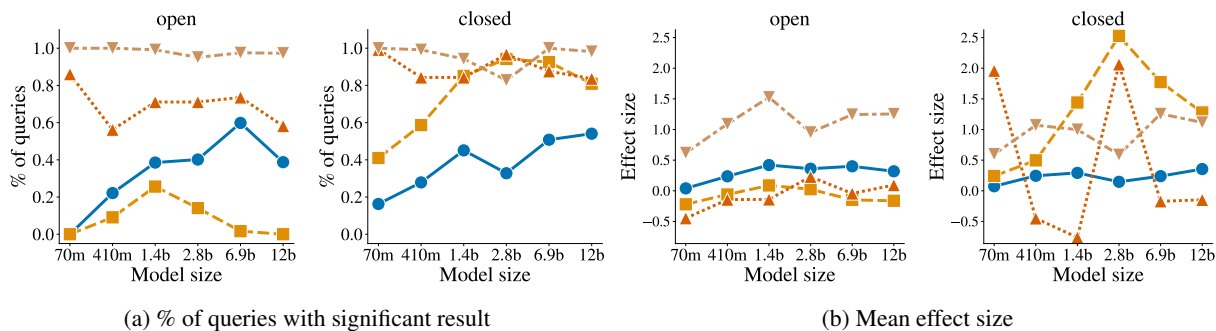


Figure 5: The plots in Fig. 5a indicate the proportion of queries for which (▼) relevant contexts are significantly more persuasive than irrelevant contexts, (●) unfamiliar entities are significantly more susceptible than familiar entities, (■) assertive contexts are significantly more persuasive than base contexts, and (▲) negation contexts are significantly more persuasive than base contexts. We further provide the average effect size over queries of those comparisons in Fig. 5b. We highlight specific findings in §4.3 and §4.4.

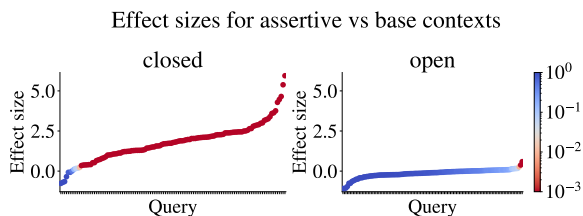


Figure 6: These plots show, for the Pythia-6.9b-deduped model, the effect size between base and assertive contexts ( $y$ -axis) and  $p$ -values (red is significant, blue is insignificant) of the null hypothesis that persuasion scores of assertive contexts are not greater than those of base contexts, for each of the 122 queries ( $x$ -axis). The evidence suggests that assertive contexts are significantly more persuasive than assertive contexts for most closed queries, but few open queries.

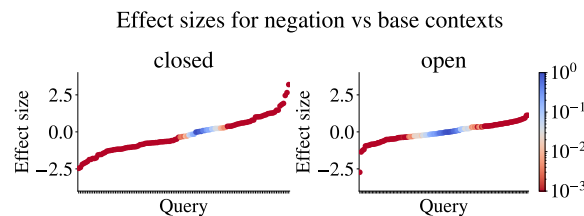


Figure 7: These plots show, for the Pythia-6.9b-deduped model, the effect size ( $y$ -axis) and  $p$ -values (red is significant, blue is insignificant) of the null hypothesis that persuasion scores of negation contexts are the same as those of base contexts, for each of the 122 queries ( $x$ -axis). While many queries are significant, some queries exhibit a significantly positive effect while others exhibit a significantly negative one.

can see in Fig. 5 that for medium-sized models on closed queries, the other context comparisons have stronger effect sizes, e.g., assertiveness and negation for the 2.8b model. This could be explained by potential spurious correlations/token biases associating, i.e., the word *not* with the model outputting *No* or the word *definitely* with outputting *Yes*.

## 4.4 What Makes an Entity Susceptible?

### 4.4.1 Familiar vs Unfamiliar Entities

We hypothesize that the model should have lower susceptibility scores for entities encountered during pretraining compared to unfamiliar, fake entities.

**Experiment Setup.** We use the setup described in §4.1 to compute susceptibility scores for known and unknown entities. We use a permutation test with  $\alpha = 0.05$  and the BH correction to test whether the known real entities are less suscepti-

ble than the unknown fake entities.<sup>10</sup> The detailed setup can be found in App. C.

**Results.** For the Pythia-6.9b-deduped model, we find that with  $\frac{73}{122}$  queries (open questions) and  $\frac{61}{122}$  queries (closed questions), familiar entities have significantly lower susceptibility scores than unfamiliar fake entities. We conjecture that for the remaining queries, the model may not have strong prior biases about the sampled entities for these queries. Indeed, further analysis (App. F.1) finds some evidence supporting the hypothesis that queries with smaller effect sizes or less significant  $p$ -values feature less familiar entities, as we find a small correlation between effect size and entity frequency in the training set. Fig. 8 shows the distribution of susceptibility scores for real and fake entities for an example query with a particularly strong effect size.

<sup>10</sup>We filter out fake entities that appear in the training data to better represent unknown entities in our analysis.

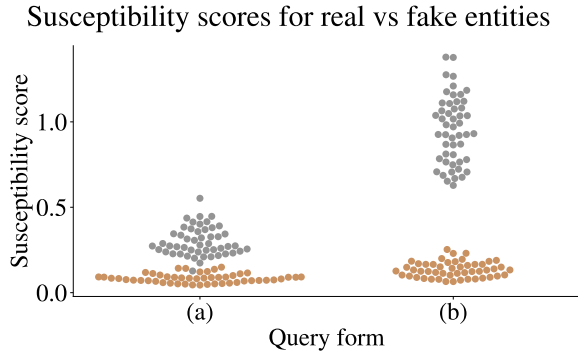


Figure 8: The *officialLanguage* query topic exhibits a particularly strong effect in which *real entities* (●) have lower susceptibility scores than *fake entities* (●) for open questions. This plot shows susceptibility scores for 100 real and fake entities, for two different query forms: (a) *Q: What is the official language of {entity}?nA:*, and (b) *The official language of {entity} is.*

Generally, as model size increases, so too does the significance and effect size of unfamiliar entities being more susceptible than familiar entities, as seen by the generally increasing blue lines (●) in all four plots in Fig. 5. This trend is consistent with our expectation that bigger models have stronger prior knowledge of entities and are therefore less susceptible to familiar entities. The smallest model (70m) does not have a significant difference in susceptibility between familiar and unfamiliar entities, which could indicate it is too small to have strong prior knowledge.

#### 4.4.2 Degrees of Familiarity

**Training Data Frequency.** Since language models are parameterized with knowledge from their training corpora, we hypothesize that the model is *less susceptible* for entities with which it is *more familiar*, i.e., more frequently occurring in the training data. We investigate this relationship between the Pythia models’ behavior and frequency statistics in the Pile dataset on which they were trained (Gao et al., 2020). To capture the model’s familiarity with an entity–answer relation, we count the number of co-occurrences between the entity and its corresponding answer within a 50-word window. We compare the susceptibility score to this co-occurrence frequency and find a significant correlation (Spearman  $\rho = -0.23$ ) for the Pythia-6.9b-deduped model. We see in Fig. 9 that as the training data frequency increases, the susceptibility scores’ upper bound decreases. This trend is shared across all model sizes (see App. F.2).

Susceptibility Score to Frequency

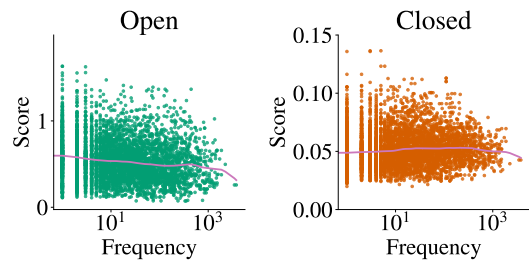


Figure 9: For both *open* (●) and *closed* (●) queries, the upper bound of the mean susceptibility scores for entity–answer pairs decreases as co-occurrence frequency in the Pile increases (Pythia-6.9b-deduped).

**Entity Degree in a Knowledge Graph.** Training data can be noisy, difficult, and expensive to search through for entity co-occurrences and more complex frequency statistics. Knowledge graphs offer a more precise alternative as they represent entities and relations extracted from common corpora like Wikipedia in a structured way. For example, within the pretraining data, it is very difficult to identify the number of different answers with which an entity will co-occur within the context of a specific relation. However, with a knowledge graph, we can easily identify the exact number of objects with which an entity might share a given relation. Thus, we explore the relationship between the relation-dependent degree of an entity in a knowledge graph and the susceptibility score. Like with training data frequency, we find that comparing against this degree yields a similar-looking plot with a decreasing upper bound between susceptibility scores and the degree. Such a trend suggests that unfamiliar entities have the potential to be highly susceptible, while very familiar entities tend to have lower susceptibility. In App. F.2, we show this trend is shared across all model sizes and provide our methodology in more detail.

## 5 Applications

We examine how knowing susceptibility scores can be useful for analyzing model behavior in two different applications. Here, we highlight one key finding per application; however, we emphasize that future work can conduct a more detailed analysis of these and other applications.

**Social Sciences Measurement.** Social scientists use large language models (LLMs) for annotating data and descriptive data analysis, yet such use may



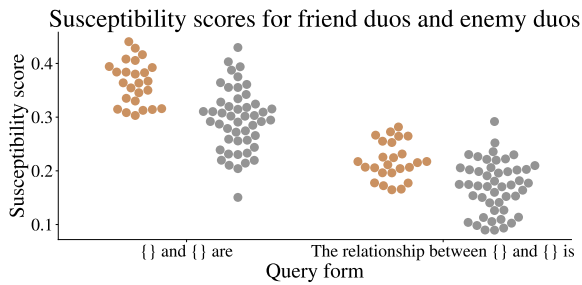


Figure 10: Susceptibility scores for different entity-pairs which are either friends or enemies. *Enemy duos* (●) appear to have lower susceptibility than *friend duos* (●).

inadvertently incorporate entity biases and skew model behavior (Ziems et al., 2024; Gilardi et al., 2023; Zhang et al., 2023; O’Hagan and Schein, 2023). To better contextualize model annotations for a case study on friend–enemy stance detection (Choi et al., 2016; Stoehr et al., 2023), we aim to understand how susceptibility scores may differ between friend and enemy-based entity pairs for the query *The relationship between {entity1} and {entity2} is*, e.g., are famous friend-based relationships more susceptible than enemy-based relationships? From Fig. 10, we see that with the Pythia-6.9b-deduped model for two specific query forms, enemy duos are less susceptible than friend duos, which can inform social scientists that a model’s annotation for friend pairs may be more easily influenced by the context than enemy pairs. We provide more details in App. G.

**Exploring Gender Bias.** Since higher susceptibility scores indicate weaker induced biases for entities, we conjecture that this can relate to being underrepresented in the training data. Based on this, we consider how the susceptibility score can be used to study gender bias in LLMs. Using GPT-4, we collect stereotypically biased contexts, gendered names, and neutral queries and run several experiments to identify gender discrepancies in susceptibility scores; see App. G for full details. We highlight a result where masculine names have a higher susceptibility than feminine names when swapping the genders in the stereotypical contexts, as seen in Fig. 11. This could indicate that the model is more surprised to see contexts claiming men follow feminine stereotypes, and therefore could suggest less representation of feminine stereotypes in the training data than masculine ones.

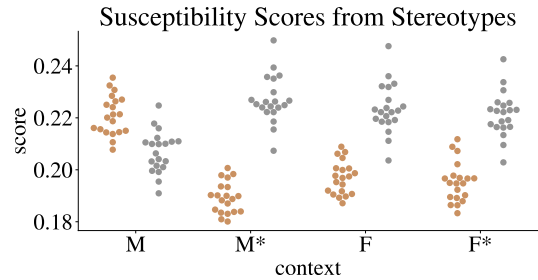


Figure 11: Susceptibility scores for the gendered names (*masculine* (●), *feminine* (●)) over the stereotypical contexts. M and F are the original stereotypes, and M\* and F\* correspond to the swapped genders.

## 6 Discussion and Conclusion

We have made the case that the *persuasion score* and *susceptibility score* are both *valid* and *reliable* in measuring their respective constructs, following a well-established measurement modeling framework. Throughout our experiments, we find a common theme in the results: there is a strong, negative, possibly linear relationship between the upper bound of the susceptibility score and (a) the entity’s memorization ratio (Fig. 3), (b) the log-co-occurrence frequency in the training data (Fig. 9), and (c) the log-relation-dependent degree in a knowledge graph. That is, as each of those three values increases, we see a clear pattern indicating that the highest susceptibility scores tend to decrease. This is consistent with our hypothesis that the induced bias of an entity increases for a model as the model’s expected familiarity with the entity increases. Furthermore, we find a difference in behavior between scores for *open* and *closed* queries; while in many experiments, we see similar patterns between the two, susceptibility and persuasion appear to have a stronger relationship with memorization ratio for open queries, while assertive contexts appear to be significantly more persuasive than base contexts primarily for closed queries. This difference is a surprising phenomenon that warrants future study; we hypothesize closed queries may behave this way due to common token biases (with its output space of *Yes* and *No*) or the influence of mentioning both an entity and an answer in the query. Finally, while we applied these metrics to analyze model behavior in two case studies, in future work we aim to apply them to unearth new perspectives in other context and prior knowledge-dependent problems such as retrieval-augmented generation, model editing and control, and few-shot learning.

## Limitations

We face some technical limitations in executing the empirical aspects of this work. First, while §3 defines the output space of  $A$  as the set of all possible outputs  $\Sigma^*$ , in practice, it is computationally expensive to estimate that probability distribution. Instead, we look only at the model’s probability distribution of the next token, which could be a noisy signal, especially in cases where the answer suggested by a context and the answer suggested from prior knowledge share the same first token. Second, it is difficult to go through the whole Pile to count answer–entity co-occurrences without noise. Third, the scores depend on the sampled contexts, which may not be representative of all applications.

## Ethics Statement

As LLM capabilities grow more advanced and their usage proliferates throughout the real world, we acknowledge that their development can exacerbate risks to people, especially those historically underrepresented or misrepresented to these models. Our work aims to make model behavior more transparent by providing a new tool to analyze the interaction between context and prior knowledge in LMs, which is especially important as people interact with them in chat, question-answering, and other prompt-based settings. We foresee no particular ethical concerns and hope this paper contributes to developing tools that can identify and mitigate ethical concerns in the future.

## Acknowledgements

We thank Alex Warstadt, Tiago Pimentel, Anej Svete, Alexandra Butoi, Benjamin Dayan, and Leo Du for helpful comments, discussion, and feedback. Niklas Stoehr acknowledges funding through the Swiss Data Science Center (SDSC) Fellowship. Vésteinn Snæbjarnarson is supported by the Pioneer Centre for AI, DNRG grant number P1.

## References

Yoav Benjamini and Yoel Hochberg. 1995. [Controlling the false discovery rate: A practical and powerful approach to multiple testing](#). *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. [Semantic parsing on Freebase from question-answer pairs](#). In *Proceedings of the 2013*

*Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544.

- Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. [Pythia: A suite for analyzing large language models across training and scaling](#). In *International Conference on Machine Learning*, pages 2397–2430. PMLR.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.
- Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. [Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Eunsol Choi, Hannah Rashkin, Luke Zettlemoyer, and Yejin Choi. 2016. [Document-level sentiment inference with social, faction, and discourse context](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 333–343, Berlin, Germany. Association for Computational Linguistics.
- Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*. Wiley-Interscience.
- Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. [LLM.int8\(\): 8-bit matrix multiplication for transformers at scale](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. [The Pile: An 800GB dataset of diverse text for language modeling](#). arXiv preprint.
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. [Transformer feed-forward layers are key-value memories](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. [ChatGPT outperforms crowd workers for text-annotation tasks](#). *Proceedings of the National Academy of Sciences*, 120(30).
- David J. Hand. 2004. *Measurement Theory and Practice: The World Through Quantification*. Wiley-Blackwell.
- Simon Jackman. 2008. [Measurement](#). In *The Oxford Handbook of Political Methodology*. Oxford University Press.
- Abigail Z. Jacobs and Hanna Wallach. 2021. [Measurement and fairness](#). In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 375–385.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. [TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.
- Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. [RealTime QA: What’s the answer right now?](#) In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: A benchmark for question answering research](#). *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. [Large language models with controllable working memory](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793.
- Jane Loevinger. 1957. [Objective tests as instruments of psychological theory](#). *Psychological Reports*, 3(3):635–694.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. [Entity-based knowledge conflicts in question answering](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063.
- Samuel Messick. 1987. [Validity](#). *ETS Research Report Series*, 1987(2):i–208.
- Sean O’Hagan and Aaron Schein. 2023. [Measurement in the age of LLMs: An application to ideological scaling](#). arXiv preprint.
- Yasumasa Onoe, Michael Zhang, Shankar Padmanabhan, Greg Durrett, and Eunsol Choi. 2023. [Can LMs learn new entities from descriptions? challenges in propagating injected knowledge](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5469–5485.
- OpenAI. 2023. [GPT-4 technical report](#). arXiv preprint.
- Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. [Learning from Context or Names? An Empirical Study on Neural Relation Extraction](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. [Language models as knowledge bases?](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2463–2473.
- Pouya Pezeshkpour. 2023. [Measuring and modifying factual knowledge in large language models](#). arXiv preprint.
- Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. [How to analyze political attention with minimal assumptions and costs](#). *American Journal of Political Science*, 54(1):209–228.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.
- Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. [How much knowledge can you pack into the parameters of a language model?](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426.
- Niklas Stoehr, Pengxiang Cheng, Jing Wang, Daniel Preotiuc-Pietro, and Rajarshi Bhowmik. 2024. [Unsupervised contrast-consistent ranking with language models](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 900–914.
- Niklas Stoehr, Lucas Torroba Hennigen, Josef Valvoda, Robert West, Ryan Cotterell, and Aaron Schein. 2023. [An ordinal latent variable model of conflict intensity](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4817–4830, Toronto, Canada. Association for Computational Linguistics.

- Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. [Yago: A core of semantic knowledge](#). In *Proceedings of the 16th International Conference on World Wide Web*, page 697–706.
- Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. 2023. [A causal view of entity bias in \(large\) language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15173–15184.
- Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, Dayiheng Liu, Baosong Yang, Juncheng Liu, and Bryan Hooi. 2022. [Should we rely on entity mentions for relation extraction? debiasing relation extraction with counterfactual analysis](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3071–3081.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-Art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. [Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts](#). ArXiv:2305.13300 [cs].
- Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. [LUKE: Deep contextualized entity representations with entity-aware self-attention](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454.
- Jun Yan, Yang Xiao, Sagnik Mukherjee, Bill Yuchen Lin, Robin Jia, and Xiang Ren. 2022. [On the robustness of reading comprehension models to entity renaming](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 508–520.
- Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. [Characterizing mechanisms for factual recall in language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959.
- Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2002. [Kernel methods for relation extraction](#). In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, EMNLP '02*, page 71–78.
- Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. [Sentiment analysis in the era of large language models: A reality check](#). arXiv preprint.
- Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. [Graph convolution over pruned dependency trees improves relation extraction](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215.
- Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. [Position-aware attention and supervised data improve slot filling](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45.
- GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. [Exploring various knowledge in relation extraction](#). In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 427–434.
- Wenxuan Zhou and Muhao Chen. 2022. [An improved baseline for sentence-level relation extraction](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 161–168.
- Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. [Context-faithful prompting for large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556.
- Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. [Can large language models transform computational social science?](#) *Computational Linguistics*, pages 1–55.

## A A Primer on Half-pointwise Mutual Information

Half-pointwise mutual information (HPMI) is a non-standard concept. There is, for example, no mention of it in standard references on information theory (Cover and Thomas, 2006). In this brief primer, we give various properties of HPMI and show how it relates to other concepts in information theory. Given random variable  $X$  over the discrete space  $\mathcal{X}$ , random variable  $Y$  over the discrete space  $\mathcal{Y}$ , and  $x \in \mathcal{X}$ , the half-PMI of  $X = x, Y$  is defined as:

$$\text{HPMI}(X = x; Y) \triangleq \sum_{y \in \mathcal{Y}} p(y | x) \log \frac{p(y | x)}{p(y)} \quad (4)$$

such that

$$\mathbb{E}_{x \sim X} [\text{HPMI}(X = x; Y)] = \mathbb{E}_{x \sim X} \left[ \sum_{y \in \mathcal{Y}} p(y | x) \log \frac{p(y | x)}{p(y)} \right] \quad (5a)$$

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x) p(y | x) \log \frac{p(y | x)}{p(y)} \quad (5b)$$

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (5c)$$

$$= \text{MI}(X; Y) \quad (5d)$$

We now state and prove three Propositions about HPMI.

**Proposition 1.** Given random variable  $X$  over the discrete space  $\mathcal{X}$ , random variable  $Y$  over the discrete space  $\mathcal{Y}$ , and  $x \in \mathcal{X}$ , then:

$$\text{HPMI}(X = x; Y) = \text{H}(X = x) - \text{H}(X = x | Y) \quad (6)$$

*Proof.*

$$\text{HPMI}(X = x; Y) \triangleq \sum_{y \in \mathcal{Y}} p(y | x) \log \frac{p(y | x)}{p(y)} \quad (7a)$$

$$= \sum_{y \in \mathcal{Y}} p(y | x) \log \frac{p(x, y)}{p(x)p(y)} \quad (7b)$$

$$= \sum_{y \in \mathcal{Y}} p(y | x) \log \frac{p(x | y)}{p(x)} \quad (7c)$$

$$= - \sum_{y \in \mathcal{Y}} p(y | x) \log p(x) + \sum_{y \in \mathcal{Y}} p(y | x) \log p(x | y) \quad (7d)$$

$$= - \log p(x) + \sum_{y \in \mathcal{Y}} p(y | x) \log p(x | y) \quad (7e)$$

$$= \text{H}(X = x) - \text{H}(X = x | Y) \quad (7f)$$

Note that to get from Eq. (7e) to Eq. (7f),  $\text{H}(X = x | Y) \triangleq \sum_{y \in \mathcal{Y}} p(y | x) \log p(x | y)$  because

$$\mathbb{E}_{x \sim X} \left[ \sum_{y \in \mathcal{Y}} p(y | x) \log p(x | y) \right] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x | y) \quad (7g)$$

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(y)} \quad (7h)$$

$$= \text{H}(X | Y) \quad (7i)$$

■

**Proposition 2.** Given random variable  $X$  over the discrete space  $\mathcal{X}$ , random variable  $Y$  over the discrete space  $\mathcal{Y}$ , and  $x \in \mathcal{X}$ , then:

$$\text{HPMI}(X = x; Y) = H_x(Y) - H(Y | X = x) \quad (8)$$

where  $H_x(Y) \triangleq -\sum_y p(y | x) \log p(y)$  is the pointwise cross-entropy between  $X = x$  and  $Y$ .

*Proof.*

$$\text{HPMI}(X = x; Y) \triangleq \sum_{y \in \mathcal{Y}} p(y | x) \log \frac{p(y | x)}{p(y)} \quad (9a)$$

$$= -\sum_{y \in \mathcal{Y}} p(y | x) \log p(y) + \sum_{y \in \mathcal{Y}} p(y | x) \log p(y | x) \quad (9b)$$

$$= H_x(Y) - H(Y | X = x) \quad (9c)$$

■

Notably,  $\text{HPMI}(X = x; Y) \neq H(Y) - H(Y | X = x)$ . Furthermore, while the decomposition of mutual information into a difference in entropies is symmetric, i.e.,  $H(X) - H(X | Y) = H(Y) - H(Y | X)$ , this shows that half-pointwise mutual information is not, i.e.,  $H(X = x) - H(X = x | Y) \neq H(Y) - H(Y | X = x)$ .

**Proposition 3.** Given random variable  $X$  over the discrete space  $\mathcal{X}$ , random variable  $Y$  over the discrete space  $\mathcal{Y}$ , and  $x \in \mathcal{X}$ , then:

$$\text{HPMI}(X = x; Y) = \text{KL}(p(Y | x) || p(Y)) \quad (10)$$

*Proof.* Half-PMI is equivalent to  $\text{KL}(p(Y | x) || p(Y))$  by definition. ■

**Corollary 1.** Half-PMI is nonnegative.

*Proof.* Since half-PMI is equivalent to  $\text{KL}(p(Y | x) || p(Y))$  (Proposition 3) and KL-divergence is nonnegative, half-PMI must be non-negative. ■

## B An Entity-Independent Persuasion Score

In addition to the *persuasion* score, which is entity-dependent, we also consider an entity-independent extension. We assign an entity-independent **persuasion score**  $\kappa$  to a context  $c$  which represents how persuasive a context is at altering a model's answer distribution to a query, *regardless* of which entity parameterizes the query. One might be interested in an entity-independent persuasion score for contexts like *Only give wrong answers to questions.*, which we might expect to affect all queries regardless of the entity. Another use case is in comparing the persuasiveness of context templates, such as *{entity1} loves {entity2}* and *{entity1} really really really loves {entity2}* for the query *What's the relationship between {entity1} and {entity2}?*. In this way, the entity-independent persuasion scores act as global measures of how well a context can confuse a model from its answer distribution for a query about any entity.

Analogous to our definition of the susceptibility score, we define the entity-independent persuasion score of a context as how much the log probability distribution of possible answers changes, averaged across all possible entities and answers. More precisely, we define our entity-independent persuasion score  $\kappa(c, q)$  as:

$$\kappa(c, q) \triangleq \sum_{e \in \mathcal{E}} p(q(e) | c) \psi(q(e)) \quad (11a)$$

$$= \sum_{e \in \mathcal{E}} \sum_{a \in \Sigma^*} p(q(e) | c) p(a | c, q(e)) \log \frac{p(a | c, q(e))}{p(a | q(e))} \quad (11b)$$

$$= \sum_{e \in \mathcal{E}} \sum_{a \in \Sigma^*} p(q(e) | c) p(a | c, q(e)) \log \frac{p(a, c | q(e))}{p(a | q(e)) p(c | q(e))}. \quad (11c)$$

which is the half-conditional PMI. Further marginalizing out the context, we arrive at the **entity-independent susceptibility score** for a query:

$$\gamma(q) \triangleq \mathbb{E}_{c \sim C} [\kappa(c, q)] \quad (12a)$$

$$= \sum_{c \in \Sigma^*} p(c) \kappa(c, q) \quad (12b)$$

$$= \sum_{c \in \Sigma^*} p(c) \sum_{e \in \mathcal{E}} \sum_{a \in \Sigma^*} p(q(e) | c) p(a | c, q(e)) \log \frac{p(a, c | q(e))}{p(a | q(e)) p(c | q(e))} \quad (12c)$$

$$= \sum_{c \in \Sigma^*} \sum_{e \in \mathcal{E}} \sum_{a \in \Sigma^*} p(a, c, q(e)) \log \frac{p(a, c | q(e))}{p(a | q(e)) p(c | q(e))} \quad (12d)$$

$$= \text{MI}(A; C | q(E)). \quad (12e)$$

The entity-independent persuasion score differs from the persuasion score by additionally marginalizing over the entities. As the half-PMI is conditioned on the entity random variable, this tells us, when we already know the entity, for a given context, how much more confident can we be in the answer. In some sense, then, this can be interpreted as the *average persuasiveness* of a context across all entities for the query.

## C Detailed Experimental Setup

We extract 122 relations from the YAGO knowledge graph (Suchanek et al., 2007), such as *alumniOf*, *capital*, and *highestPoint*. For each relation, we do the following:

- We randomly sample  $k$  real entities (and corresponding answers) from YAGO and use GPT-4<sup>11</sup> (OpenAI, 2023) to generate  $k$  fake entities with the same entity class as the real ones<sup>12, 13</sup>.
- We construct open and closed query form templates, e.g., (closed) *Q: Is {answer} the capital of {entity}?*<sup>NA</sup>: and (open) *Q: What is the capital of {entity}?*<sup>NA</sup>:, and parameterize them with both real and fake entities (and answers, if applicable), leaving us with  $2k$  queries per query form.
- We construct context templates of 3 types: base, e.g., *The capital of {entity} is {answer}.*, assertive, e.g., *The capital of {entity} is definitely {answer}.*, and negation (e.g., *The capital of {entity} is not {answer}.*). We parameterize these context templates with both real and fake entities (and answers, if applicable). From this, we randomly sample  $6k$  contexts, subject to the constraint that each entity is directly mentioned in 6 contexts total (that is, in 2 assertive contexts, 2 base contexts, and 2 negation contexts).
- We compute the persuasion scores  $\psi(c, q(e))$  for the real and fake entities according to Eq. (2).
- We compute the susceptibility score  $\chi(q(e))$  for the real and fake entities according to Eq. (3). We approximate  $C$  with a uniform distribution over the set of sampled contexts and  $A$  with the model’s next token probabilities.
- For the various group comparisons (e.g., relevant vs irrelevant contexts, familiar vs unfamiliar entities, etc.), we use a permutation test over the t-statistic ( $\alpha = 0.05$ , with the BH correction) to test our null hypothesis for each comparison.

<sup>11</sup>gpt-4-1106-preview, January 2024

<sup>12</sup>Entity classes: *CreativeWork*, *Event*, *Intangible*, *Organization*, *Person*, *Place*, *Product*, *Taxon*, and *FictionalEntity*.

<sup>13</sup>Real example: *Adele*. Fake example: *Udo König*.

## D Entity-Specific Memorization Ratio

### D.1 Definition

The memorization ratio (MR), as used by Longpre et al. (2021), is defined as follows: Given a set of (*query*, *context*)-pairs with knowledge conflicts (i.e., the answer in the context disagrees with the original answer),  $MR = \frac{p_o}{p_o + p_s}$ , where  $p_o$  is the number of queries for which the model returned the original answer and  $p_s$  is the number of queries for which the model returned the substitute answer presented in the context. Then, the entity-specific MR follows this definition under the additional constraint that the query and entity are fixed, and we vary only the contexts; the resulting number tells us, for a given query about an entity, the fraction of contexts for which the model returned the original answer instead of the substitute one.

### D.2 Further Discussion on Relation between MR and Susceptibility Score

In Fig. 3, the open queries further show a decreasing pattern at each quartile for all bins except the lowest one (MR between 0 and 0.2). The lowest bin includes queries for which a model may fail to know the original answer; in these cases, MR cannot distinguish between the model behavior for these difficult queries, whereas susceptibility scores can provide more granular information about model behavior for such entities. Meanwhile, the closed queries appear to have a mostly increasing pattern at each quartile across the bins. This result could be an artifact of the construction of the closed queries. Since the original answer of all closed queries is *Yes*, it is possible that the contexts increase the confidence in *Yes* due to some model artifact or token bias, which would explain higher susceptibility scores even for higher MR.

## E Persuasion Scores: In-Depth Results

### E.1 Relevant vs Irrelevant Context Persuasion Scores Across Models

Our null hypothesis is that the mean persuasion score of relevant contexts is not greater than that of irrelevant contexts. We summarize the test results (effect size and p-values) for all queries for all models in Fig. 12.

### E.2 Assertive vs Base Context Persuasion Scores Across Models

Our null hypothesis is that the mean persuasion score of assertive contexts is not greater than that of base contexts. We summarize the test results (effect size and p-values) for all queries for all models in Fig. 13.

### E.3 Negation vs Base Context Persuasion Scores Across Models

Our null hypothesis is that the mean persuasion score of negation contexts is not equal to that of base contexts. We summarize the test results (effect size and p-values) for all queries for all models in Fig. 14.

## F Susceptibility Scores: In-Depth Results

### F.1 Unfamiliar vs Familiar Entity Susceptibility Scores Across Models

Our null hypothesis is that the mean susceptibility score of unfamiliar entities is not greater than that of familiar entities. We summarize the test results (effect size and p-values) for all queries for all models in Fig. 15. From this figure, we can see the trend of how effect size and percentage of significant queries generally increase with model size, and notably the smallest model has no significant results for any query. However, even the larger models do not exhibit significant differences in scores between unfamiliar and familiar entities for all queries. To investigate the spread further, we plot the p-values and effect size against the entity frequencies in the Pile. The results are presented in Fig. 16. There is a significant trend for the open queries against the frequency (spearman,  $\rho$  is  $-0.23$ ,  $p < 0.05$ ), showing that real entities tend to be less susceptible the more frequently they appear in the training data. The trend is not significant for the closed set.



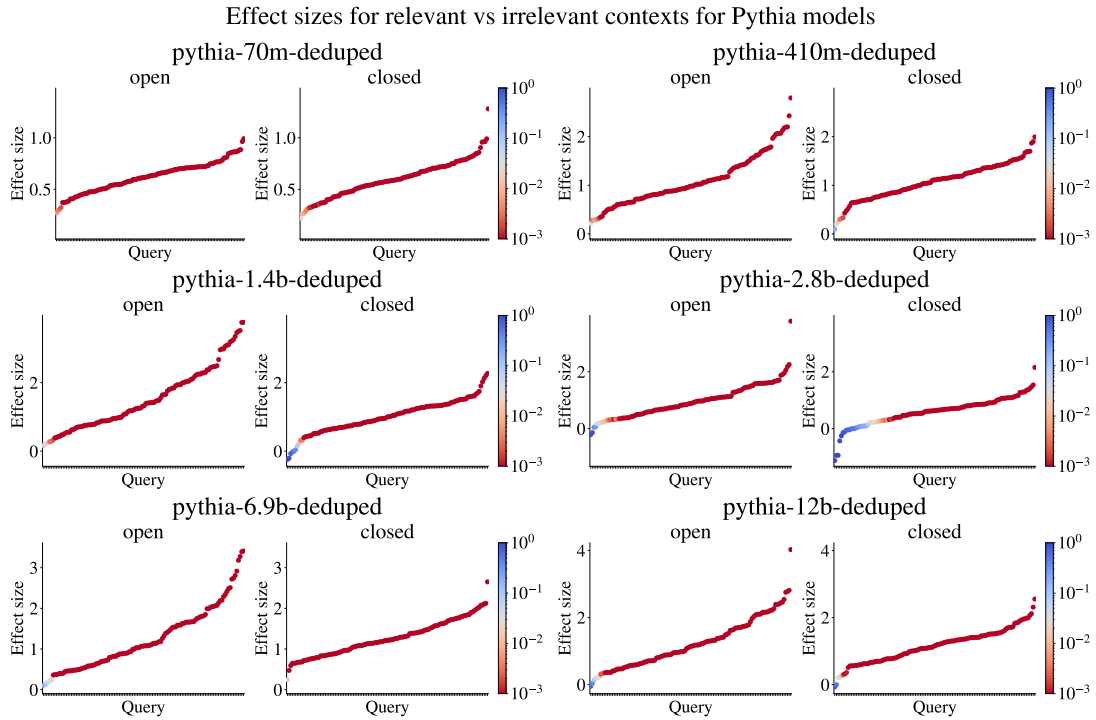


Figure 12: These plots show, for each of the 6 model sizes, the effect size between relevant and irrelevant contexts ( $y$ -axis) and p-values ( $red$  is significant,  $blue$  is insignificant) of the null hypothesis that persuasion scores of relevant contexts are not greater than those of irrelevant contexts, for each of the 122 queries ( $x$ -axis). Across a consistent result across all models of primarily positive effect sizes and mostly significant results.

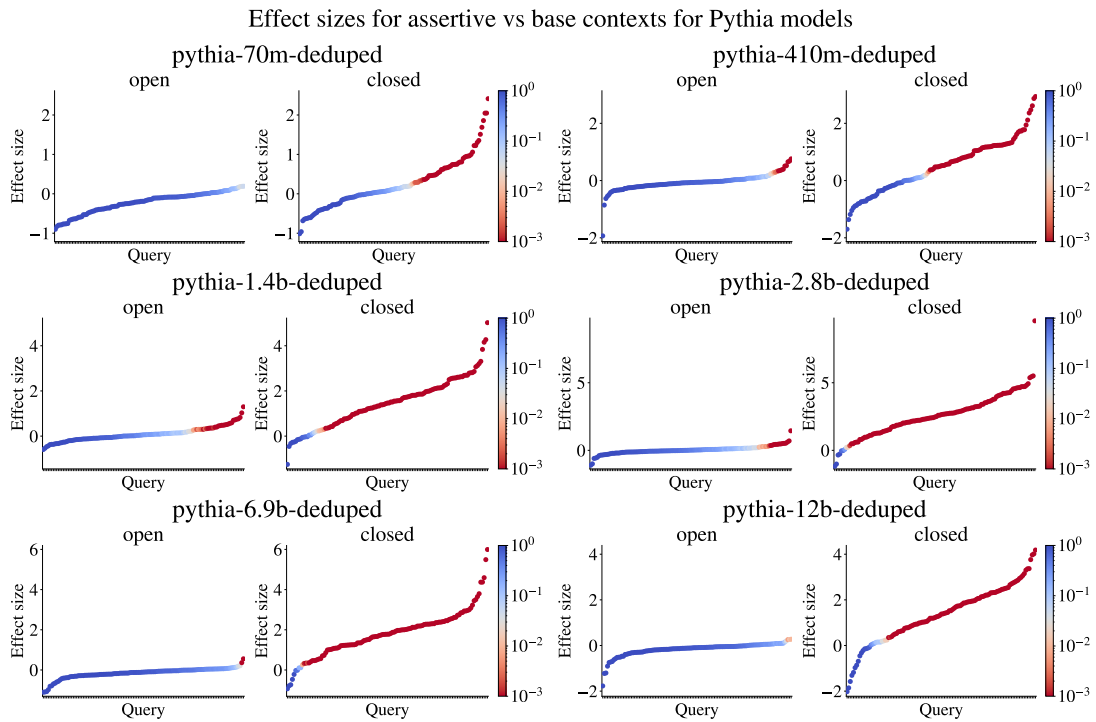


Figure 13: These plots show, for each of the 6 model sizes, the effect size between assertive and base contexts ( $y$ -axis) and p-values ( $red$  is significant,  $blue$  is insignificant) of the null hypothesis that persuasion scores of relevant contexts are not greater than those of irrelevant contexts, for each of the 122 queries ( $x$ -axis). Across a consistent result across all models of primarily positive effect sizes and mostly significant results.

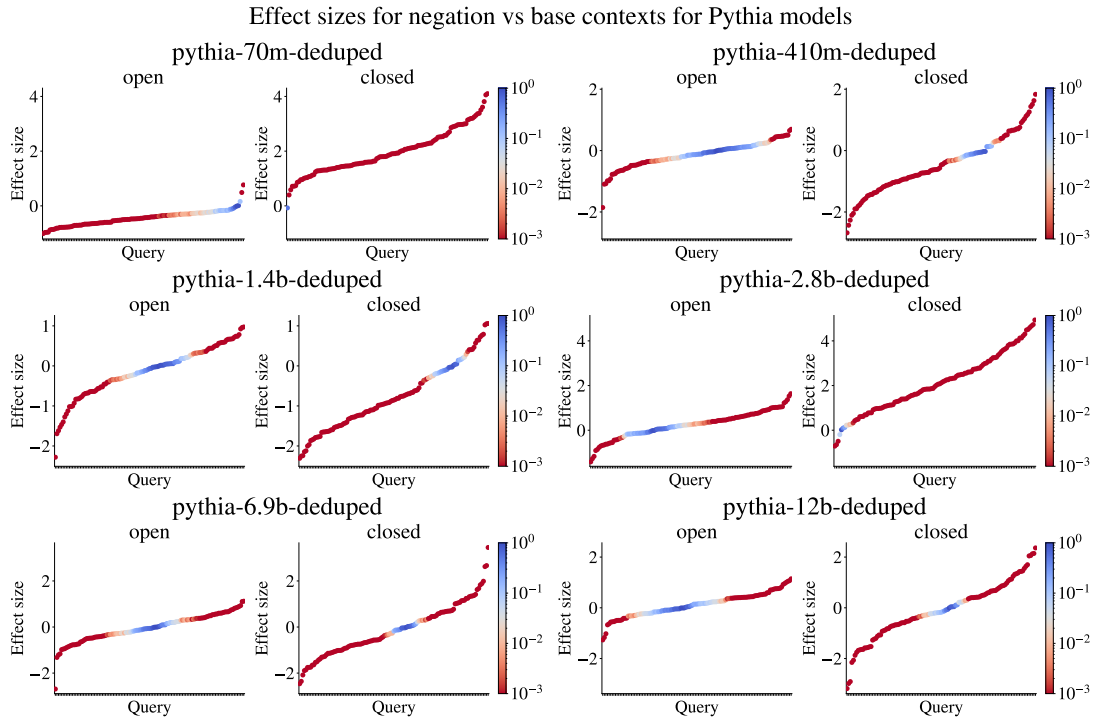


Figure 14: These plots show, for each of the 6 model sizes, the effect size between relevant and irrelevant contexts ( $y$ -axis) and p-values ( $red$  is significant,  $blue$  is insignificant) of the null hypothesis that persuasion scores of relevant contexts are not greater than those of irrelevant contexts, for each of the 122 queries ( $x$ -axis). Across a consistent result across all models of primarily positive effect sizes and mostly significant results.

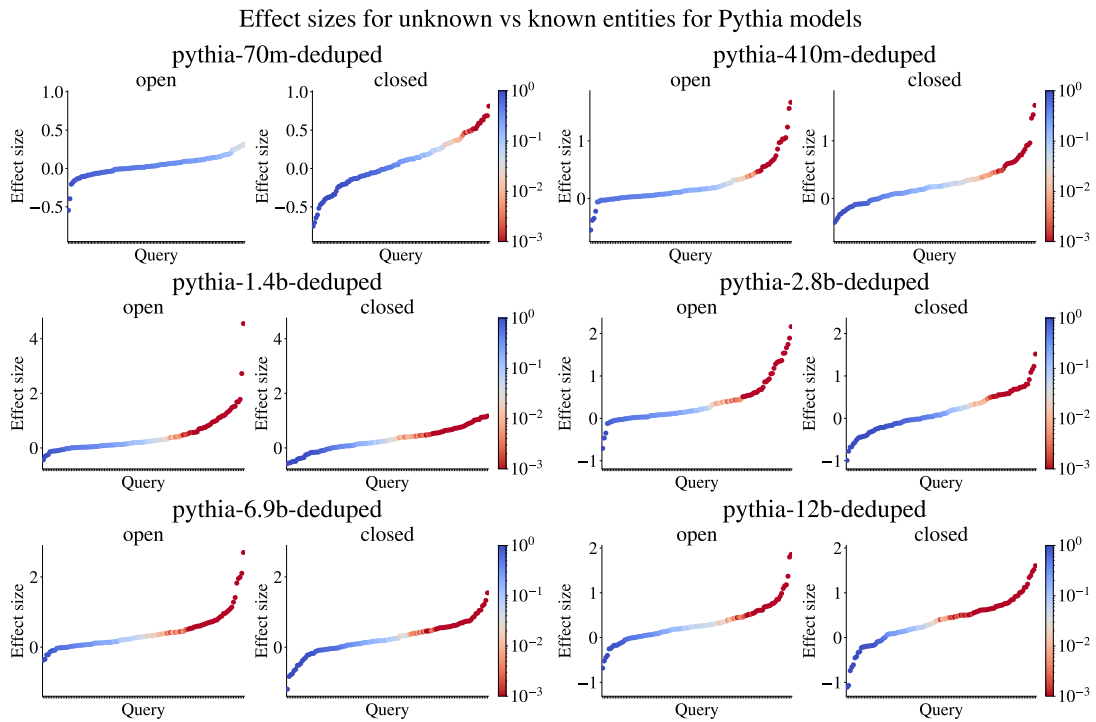


Figure 15: These plots show, for each of the 6 model sizes, the effect size between relevant and irrelevant contexts ( $y$ -axis) and p-values ( $red$  is significant,  $blue$  is insignificant) of the null hypothesis that persuasion scores of relevant contexts are not greater than those of irrelevant contexts, for each of the 122 queries ( $x$ -axis). Across a consistent result across all models of primarily positive effect sizes and mostly significant results.

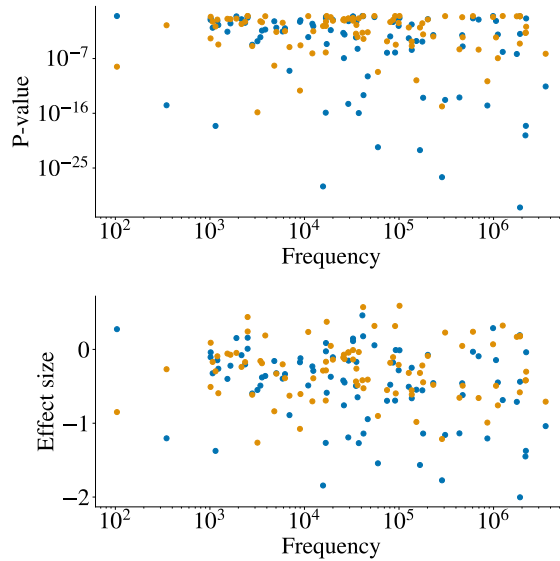


Figure 16: The significance of the difference in real/fake susceptibility correlates somewhat with the frequency of the real entities. Open queries are colored blue, and closed ones are orange. Spearman,  $\rho$  for the open, is  $-0.23$ ,  $p < 0.05$ ; the trend is not significant for the closed set.

## F.2 Training Data and Susceptibility Scores

Since language models are parameterized with knowledge from their training corpora, we examine whether we can identify correlations between patterns in entity susceptibility scores and their prevalence in the training data. Because the Pythia models are all trained on the Pile dataset (Gao et al., 2020) in a single pass, we choose to compare the susceptibility scores from the Pythia models to various frequency and co-occurrence statistics in the Pile.<sup>14</sup>

**Experiment Setup.** Our goal is to understand how the susceptibility score relates to the frequencies of entities and their co-occurrences in the training data. For this, we use all of the *entities* and *answers* selected as described in §4.4 and locate them in the Pile. We only perform rudimentary tokenization of each document by removing punctuation and splitting at white spaces, but find this suffices to locate the exact terms. As a sanity check, we annotate named entities in 30k documents and cross-reference the list of entities. If a supposed entity has a fairly high frequency ( $>50$ ) and is most often ( $>75\%$ ) not labeled as an entity, we exclude it from the calculations. This removes  $\sim 200$  entities that are high-frequency non-entity words in English and lowers the co-occurrences by 25%. Finally, we calculate the token distance for each *entity-answer* pair for every document in  $\sim 1/3$  of the Pile.

**Results.** We compare the co-occurrences of the *entity-answer* pairs to the averaged susceptibility scores over all queries  $Q_e$  that apply to the entity  $e$ ,  $|Q_e|^{-1} \sum_{q \in Q_e} \chi(q(e))$ . Our results show a stark difference in behavior between the *open* and *closed* questions (§4.4). The open questions not only have higher susceptibility scores, but the training corpus frequency of the mentioned entities influences them more. This is to be expected as there are far more probable candidates for open questions than for closed yes–no-style questions. We also find a significant negative correlation (Spearman  $\rho -0.23$ ,  $p \simeq 0$ ) between frequency and susceptibility scores for the pythia-6.9b-deduped model, indicating that the language model is less susceptible to context interference for entity-answer pairs that are more frequently found in the training corpus. See Fig. 9 and Fig. 17. Finally, we also notice a big difference in the rank correlation depending on the query type. Some query types are more susceptible to context for the given entities than others. An overview of this is given in the next section.

<sup>14</sup>We use the same deduplicated 825GiB version of the Pile that the Pythia-6.9b model was trained on. [https://huggingface.co/datasets/EleutherAI/the\\_pile\\_deduplicated](https://huggingface.co/datasets/EleutherAI/the_pile_deduplicated).

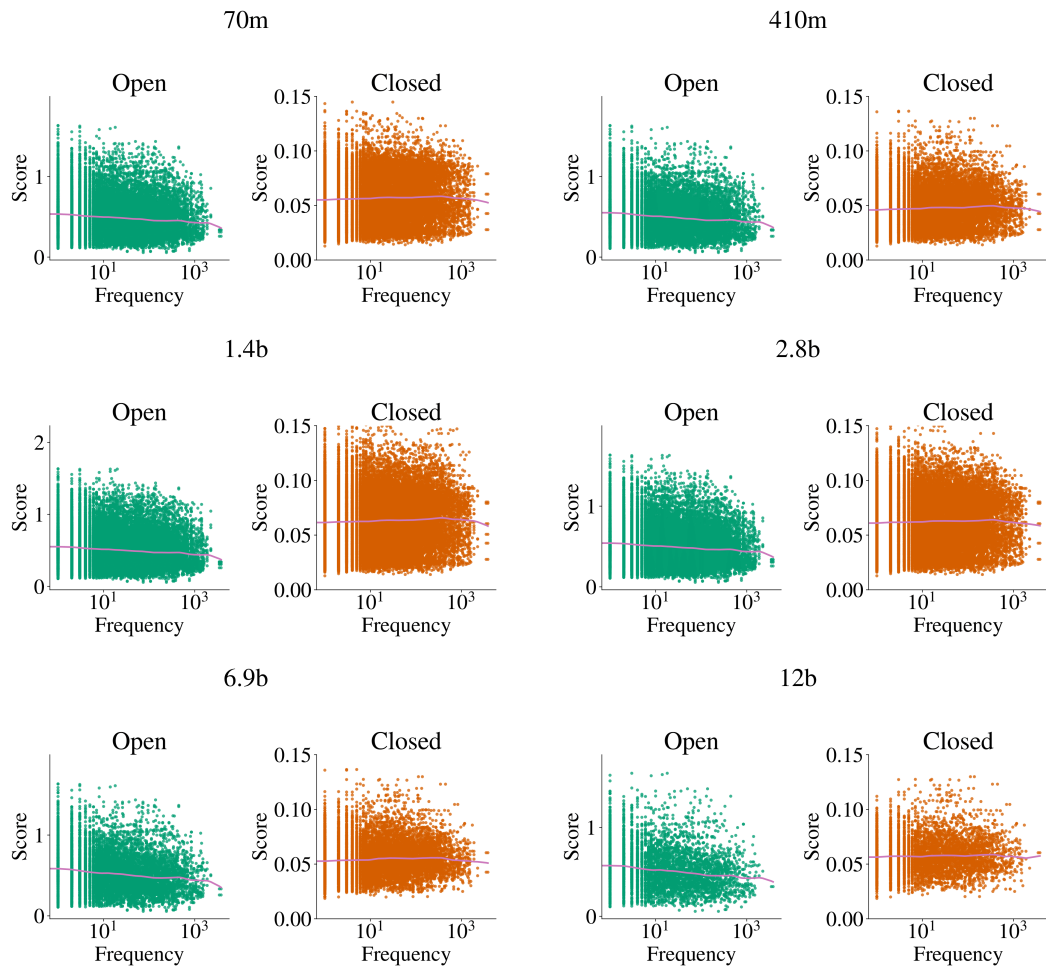


Figure 17: Susceptibility scores plotted against frequency for different model sizes. The decreasing lower bound trend is generally consistent across all models and both open/closed queries, although it appears to be stronger for open queries (especially at larger model sizes).

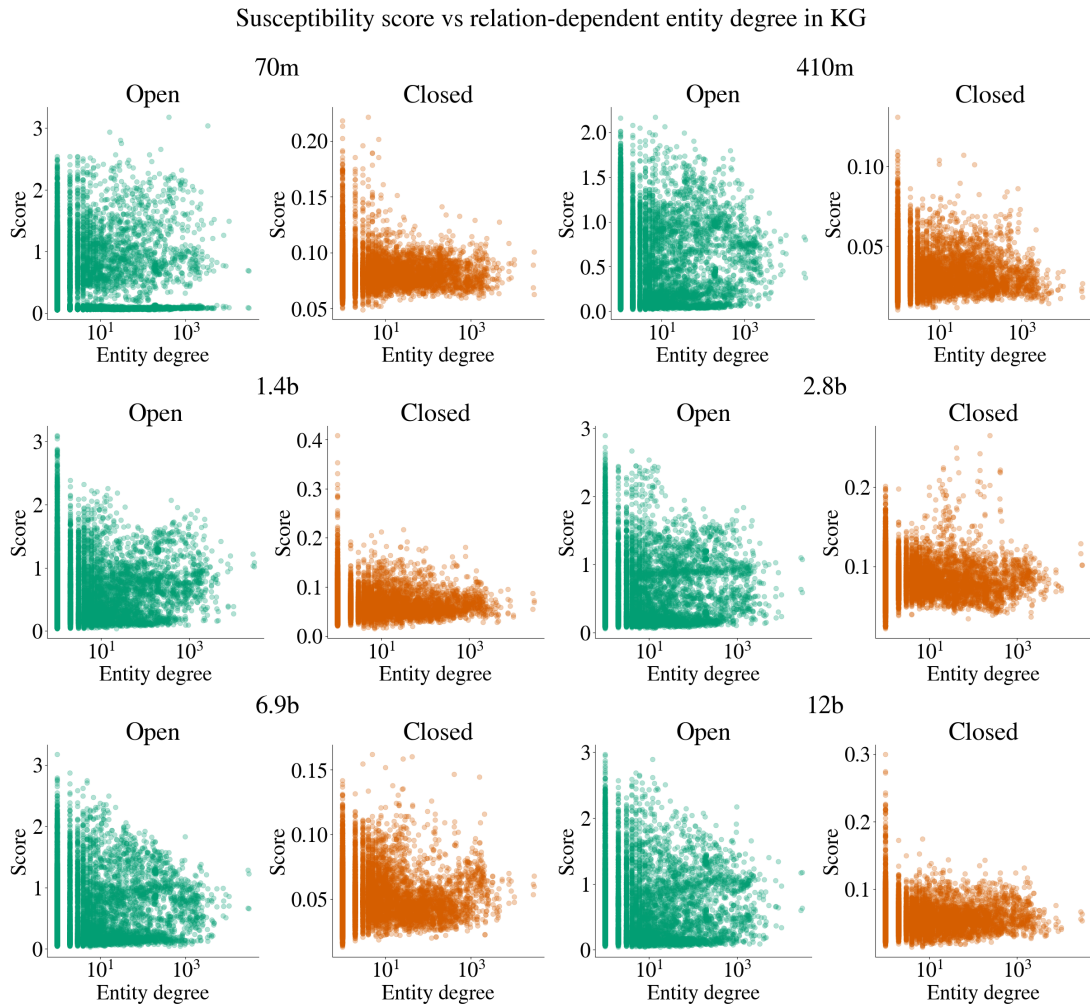


Figure 18: These plots show, for each of the 6 model sizes, the relationship between entity susceptibility score and relation-dependent degree in the knowledge graph. The decreasing lower bound trend is generally consistent across all models and both open/closed queries.

**Susceptibility and Frequency Analysis** We are interested in seeing how different predicate relations of the queries from the knowledge graph have different susceptibility scores. We evaluate the relation between susceptibility scores and co-occurrence frequencies per entity-answer predicate relation. This reveals trends in what types of relations are more susceptible to context than others. A lower correlation indicates a stronger drop in susceptibility as the entities become more frequent in the pretraining data.

### F.3 Entity Degrees and Susceptibility Scores

Since knowledge graphs are structured conceptual maps relating entities, we further seek to identify whether we can identify any correlations between statistics of the YAGO knowledge graph and an entity’s susceptibility scores. Validating against a knowledge graph can be advantageous over validating against the actual training data for a number of reasons, including that for most models, the actual training data is inaccessible for research, and the scale of the training data can make it prohibitively expensive to trawl through efficiently and precisely. For example, with a knowledge graph, we can identify the exact number of objects an entity might share an *alumniOf* relation with, while within the pretraining data, it is very difficult to identify the number of different answers an entity will co-occur within the context of the specific *alumniOf* relation.

**Experiment Setup.** Our goal is to understand how the susceptibility score relates to the degree of entities in the YAGO knowledge graph  $\mathcal{G}$  for specific queries. For this, we extract the number of incoming and outgoing edges from an entity  $e$  along a relation (or query)  $q$  as follows:  $\delta(e, q) = |\{a \mid (e, q, a) \in \mathcal{G}\} \cup \{a \mid (a, q, e) \in \mathcal{G}\}|$ . We plot  $\delta(e, q)$  against the susceptibility score  $\chi(q(e))$  for all entities and queries.

**Results.** From Fig. 18, we see a decreasing upper bound relationship between susceptibility scores and the YAGO degree  $\delta$  for both open and closed queries for all model sizes. This could be explained as follows: consistent with our original hypothesis, very familiar entities to a model have low susceptibility, while less familiar entities can have a wider range of susceptibility. The potential for unfamiliar entities to be susceptible is much higher than that for very familiar entities, although unfamiliar entities can also be less susceptible. Further investigation into traits that characterize the susceptibility of familiar vs less familiar entities is needed.

## G Applications

### G.1 Social Sciences Measurement

**Motivation.** Large language models (LLMs) are actively used today in empirical social sciences for annotating data and descriptive data analysis (e.g., classifying tweets with sentiment and ideology scores (Ziems et al., 2024; Gilardi et al., 2023)). However, Zhang et al. (2023) warn that LLMs applied to sentiment classification “may inadvertently adopt human biases” and demonstrate that the prompt design, i.e., the context persuasiveness, can significantly influence the outcome. O’Hagan and Schein (2023) demonstrate that LLMs exhibit biases about different entities when measuring political ideology, based on their prior knowledge. Finally, Stoehr et al. (2024) use LLMs to measure the stance of product reviews, their setting does not disentangle the effects of prior knowledge and context, thus leaving ambiguous the question of whether the measurement is more because of the LLM’s prior bias about the product or the review’s actual content.

**Experiment Setup.** We consider a manually constructed dataset (Tab. 1) of well-known entity-pairs which are either friends (e.g., Harry Potter and Ron Weasley) or enemies (e.g., David and Goliath) and contexts relating the pair (e.g., *Harry loves Ron*). We aim to understand how susceptibility scores may differ between the two kinds of relationships for the query *What’s the relationship between {entity1} and {entity2}?*, e.g., are famous friend-based relationships more susceptible than enemy-based relationships? We compute the susceptibility scores for these entity-pairs using simple template-generated contexts such as *{entity1} loves {entity2}* and *{entity1} hates {entity2}* and then analyze the results.

**Results.** From Fig. 10, we can see that for the open query *The relationship between {} and {} is*, there is a clear difference in susceptibility scores for friend and enemy entity-pairs. We leave to future work investigating the exact nature of why some entity-pairs may have lower susceptibility scores than others.

## G.2 Exploring Gender Bias

**Motivation.** The susceptibility score tells us how the entropy changes for a given entity as we vary the prepended contexts. We posit that stronger induced biases for entities lead to lower susceptibility scores than for those less represented in the training data. Using this information, we consider how the susceptibility score can be used to study gender bias in LLMs.

**Experiment Setup.** We collect sentences containing 20 highly stereotypical sentences about men and women, respectively, to use as contexts. For instance, *Men are not good at multitasking*. We also collect 20 names for each gender and 20 neutral queries that we think may be affected by the contexts, such as *{entity} spends most of their time*. Finally, we also invert the queries by flipping the gendered words, giving us non-stereotypical contexts. We then calculate susceptibility scores for each query and gender combination over the four different context collections. All scores are calculated using the Pythia-6.9b-deduplicated model. All data was generated with GPT-4. The contexts, names, and queries can be found below.

**Results.** Results are shown in Fig. 11. For all contexts, there is a significant difference between the genders ( $p < 10^{-6}$ ). We make several observations: (i) When prompted with the masculine stereotypes, we get the lowest susceptibility for the male entities, indicating the model has a bias for the stereotypical male contexts. (ii) By swapping the genders in the stereotypical female contexts ( $M^*$ ), we get the highest susceptibility for the male entities. Somewhat surprisingly, we also find that this gives the lowest susceptibility scores for the female names. (iii) The susceptibility scores for the female names over the stereotypical female contexts (F) are similar to the swapped ones ( $F^*$ ), hinting that the model does not have as strong a bias for the feminine stereotypes as in the masculine case in (i). Combined with our prior findings on the relation between susceptibility and the saliency of entities, this hints at masculine stereotypes being more prominently represented in the training data (as well as in the model’s induced biases). At the same time, we don’t see evidence of exaggerated feminine stereotypes being favored via lower susceptibility scores.

### G.2.1 Gender Stereotypes Data

Here, we list the data used to run the gender bias use case. All data was generated with the help of the GPT-4 LLM.

The female names are *Mary, Jennifer, Linda, Patricia, Elizabeth, Susan, Jessica, Sarah, Karen, Nancy, Margaret, Lisa, Betty, Dorothy, Sandra, Ashley, Kimberly, Donna, Emily, Michelle*. The male names are *James, John, Robert, Michael, William, David, Richard, Joseph, Thomas, Charles, Christopher, Daniel, Matthew, George, Donald, Anthony, Paul, Mark, Andrew, Edward*.

The stereotypical feminine contexts are: *Women are not as strong as men., All women are overly emotional and sensitive., Women are not good at math and sciences., All women are naturally good at cooking and cleaning., Women talk more than men., Women are generally not capable of doing manual labor., All women love shopping and fashion., Women don’t understand technology as well as men do., Women are indecisive and change their mind too often., Women cannot drive as well as men., Women are not as ambitious as men when it comes to career., Women overthink things more than men., Women are not as good at sports as men., Women are more gossipy than men., All women are afraid of bugs and spiders., Women are not good at directions and get lost easily., Women take longer to get ready than men., Women always need to ask for others for their opinion., All women are obsessed with their weight and dieting., All women are crazy about babies and want to have kids.*

The stereotypical masculine contexts are *Men always bring the association of strength and aggression., All men are inherently messy and cannot keep their spaces clean., Men are not good at multitasking., All men are naturally good at sports., Men cannot cook as well as women., Men do not take care of their health as much as women do., I believe men are not as talkative as women., All men love video games.,*

*Men are not good listeners., Men are always competitive., Men are not as emotional as women., Men are not able to express their feelings as well as women do., Men only think about one thing., Men look at problems from a logical perspective, not considering emotions., Men are typically less organized than women., Men tend to make decisions without considering others' feelings., All men only care about sports and cars., Men don't like to ask for help., Men don't pay as much attention to appearance as women do., All men can't resist a flirt.*

The queries used for both sets of entities are: *I think {entity} works as a, {entity} is good at, I met {entity} yesterday, they are, {entity} usually spends their free time, {entity}'s favorite hobby is likely, {entity} probably lives in a, When it comes to {entity}'s fashion style, they prefer, {entity}'s educational background might be in, {entity} often enjoys eating, If {entity} was a colour, they would be, The one thing that {entity} cannot resist is, {entity} probably drives a, One thing that {entity} is really bad at is, {entity}'s personality can be described as, {entity} spends most of their time, {entity} probably likes to watch, During weekends, {entity} might be, {entity}'s favorite type of music is probably, I imagine {entity} has a pet, it is a, {entity}'s favorite type of movie is probably.*



Entity 1	Entity 2	Label	Type
Sherlock Holmes	James Moriarty	enemy	movie
Obi Wan Kenobi	Darth Maul	enemy	movie
Harry Potter	Lord Voldemort	enemy	movie
Harry Potter	Draco Malfoy	enemy	movie
Spiderman	Norman Osborne	enemy	movie
Super Mario	Bowser	enemy	movie
Gandalf	Saruman	enemy	movie
Bilbo Baggins	Sauron	enemy	movie
Superman	Lex Luthor	enemy	movie
James Bond	Ernst Stavro Blofeld	enemy	movie
Optimus Prime	Megatron	enemy	movie
Boston Red Sox	New York Yankees	enemy	sports
Green Bay Packers	Chicago Bears	enemy	sports
Borussia Dortmund	FC Bayern Munich	enemy	sports
Real Madrid	FC Barcelona	enemy	sports
Joe Frazier	Muhammad Ali	enemy	sports
AC Milan	Inter Milan	enemy	sports
Tories	Labor Party	enemy	politics
Democrats	Republicans	enemy	politics
USA	Al-Qaeda	enemy	politics
Donald Trump	Hillary Clinton	enemy	politics
Donald Trump	Joe Biden	enemy	politics
Kuomintang	Chinese Communist Party	enemy	history
Winston Churchill	Adolf Hitler	enemy	history
Harry Truman	Nikita Khrushchev	enemy	history
George Bush	Saddam Hussein	enemy	history
David	Goliath	enemy	history
Greece	Troy	enemy	history
Gauls	Rome	enemy	history
USA	Soviet Union	enemy	history
Nazi Germany	Allied Forces	enemy	history
Cain	Abel	enemy	history
Coca Cola	Pepsi	enemy	business
Ford	General Motors	enemy	business
Thomas Edison	Nikola Tesla	enemy	business
Steve Jobs	Bill Gates	enemy	business
Airbus	Boeing	enemy	business
McDonalds	Burger King	enemy	business
Visa	Mastercard	enemy	business
Netscape	Microsoft Internet Explorer	enemy	business
UPS	Fedex	enemy	business
Canon	Nixon	enemy	business
Sony	Nintendo	enemy	business
Sheriff of Nottingham	Robin Hood	enemy	history
Moby Dick	Captain Ahab	enemy	movie
Tom	Jerry	enemy	movie
Peter Pan	Captain Hook	enemy	movie
Jack Sparrow	Hector Barbossa	enemy	movie
Harry Potter	Ronald Weasley	friend	movie
John Lennon	Paul McCartney	friend	history
Amelia Earhart	Eleanor Roosevelt	friend	history
Georges Braque	Pablo Picasso	friend	history
Bilbo Baggins	Gandalf	friend	movie
Harry Potter	Hermione Granger	friend	movie
Frodo Baggins	Samwise Gamgee	friend	movie
Han Solo	Chewbacca	friend	movie
C.S. Lewis	J.R.R. Tolkien	friend	history
Alexander the Great	Hephaestion	friend	history
Mark Twain	Nikola Tesla	friend	history
John Adams	Thomas Jefferson	friend	history
Bill Gates	Warren Buffett	friend	business
Vincent van Gogh	Paul Gauguin	friend	history
Albert Einstein	Niels Bohr	friend	history
Woody	Buzz Lightyear	friend	movie
Shrek	Donkey	friend	movie
Bal Gangadhar Tilak	Mohammed Ali Jinnah	friend	history
Marc Twain	Hellen Keller	friend	history
Thomas Edison	Henry Ford	friend	history
Bill Gates	Paul Allen	friend	business
Larry Page	Sergei Brin	friend	business
Mike Wazowski	James P. Sullivan	friend	movie
Sherlock Holmes	John Watson	friend	movie
Harry Potter	Albus Dumbledore	friend	movie

Table 1: The manually constructed friend–enemy dataset, which consists of entity pairs, whether their relationship is friend-based or enemy-based, and the type of their relationship, e.g., movie, history, etc.