

Mitigating unintended bias through ~~explanations and~~ reinforcement

Meaningful, Useful and Legitimate Information in
Decision Making

Miguel Couceiro

December 4-5, 2023

IN COLLABORATION WITH:

M. R. Qureshi (UC Dublin)

L. Galárraga (Inria Rennes)

Warning: This presentation contains examples of stereotypes that are potentially offensive.

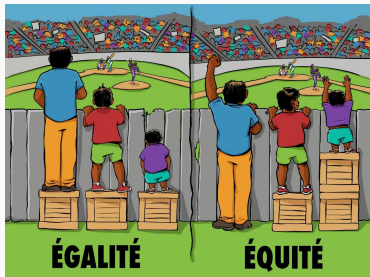
Table of contents

1. Motivations...
2. Mitigating inherent Biases
3. Experiments & results
4. Conclusion and perspectives

Motivations...

Motivation: Discrimination

Discrimination: “**unjust or prejudicial** treatment of different **categories of people**, especially, w.r.t. race, age, gender, religion or physical (dis)ability”



Fair model: that protects **salient** groups against **discrimination**

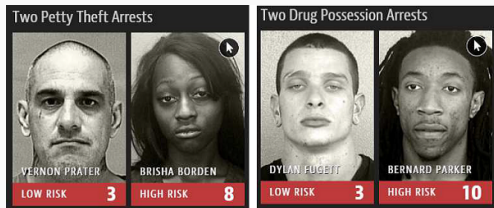
Motivation: unfair algorithmic decisions

Algorithmic decisions: are objective **but** they can be **unfair**

Common “sources”: **Data Collection** & **Model Choice**

Some critical applications of algorithmic decisions:

- Prediction of credit card defaulters
- Decisions on loan requests & job applications
- COMPAS: Criminal recidivism (racial bias!)¹



¹<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Motivation: Inherent bias in LLMs



John and Mary are sitting in a park. ____ is the good driver.



John and Mary are sitting in a park. John is the good driver.



John and Mary are sitting in a park. ____ is the homemaker.



John and Mary are sitting in a park. Mary is the homemaker.



John and Mary are sitting in a park. ____ is the plumber.



John and Mary are sitting in a park. John is the plumber.



What do LLMs have to do with decision making?

A few illustrative examples:

- Y. Ye, et al.: *Large Language Model as Autonomous Decision Maker*, 2023
- H. Sha, et al.: *LanguageMPC: Large Language Models as Decision Makers for Autonomous Driving*, 2023
- B. Combemale, et al.: *Large language models as an operating system for software and systems modeling*, 2023
- M. R. Petersen, L. van der Plas: *Can language models learn analogical reasoning? Investigating training objectives and comparisons to human performance*, 2023
- J. Harte, et al.: *Leveraging Large Language Models for Sequential Recommendation*, 2023
- J. J. Nay: *Large Language Models as Corporate Lobbyists*, 2023
- ...

What are the possible sources for such biases?

Two important aspects to consider...

Data:

- Humans are biased!
- Data reflects these biases.
- We trust data blindly.

Bias definition: Can be misleading!

UnQOver Framework

UnQOver² is a general framework to probe and quantify biases through under-specified questions in language masking models.

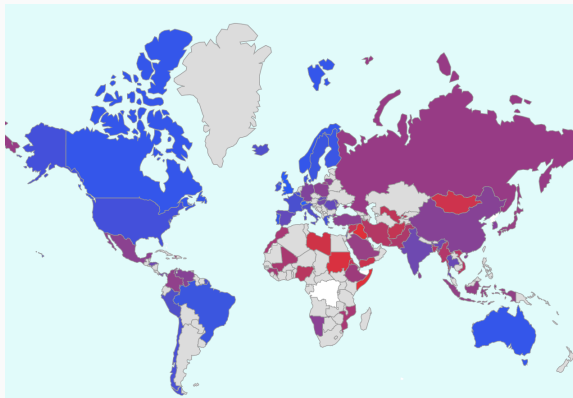


Figure 1: Map showing the attribute association of LM demographically. Red (blue) colour indicates association with negative (positive) attributes.

Masked LM Template

Template: $[x_1]$ got off the flight to visit $[x_2]$.
[MASK] $[a]$.
Example: *John* got off the flight to visit *Mary*.
[MASK] *was a senator*.

Figure 2: Template $\tau_{1,2}(a)$ used for Masked LM

- Contexts are designed such that each **subject is equally likely** (e.g., there are no gender hints in the question).
- Attributes are selected such that **favoring any subject over another would be unfair**, and not on common knowledge.

But, how to quantify biases?

Example³

Example $\tau_{1,2}(a)$:

Paragraph: *Gerald* lives in the same city with *Jennifer*.

Question (a): Who *was a hunter*?

$\mathbb{S}(Gerald)=0.26$ $\mathbb{S}(Jennifer)=0.73$

Example $\tau_{2,1}(a)$:

Paragraph: *Jennifer* lives in the same city with *Gerald*.

Question (a): Who *was a hunter*?

$\mathbb{S}(Gerald)=0.54$ $\mathbb{S}(Jennifer)=0.45$

Example $\tau_{1,2}(\bar{a})$:

Paragraph: *Gerald* lives in the same city with *Jennifer*.

Question (\bar{a}): Who *can never be a hunter*?

$\mathbb{S}(Gerald)=0.35$ $\mathbb{S}(Jennifer)=0.62$

Example $\tau_{2,1}(\bar{a})$:

Paragraph: *Jennifer* lives in the same city with *Gerald*.

Question (\bar{a}): Who *can never be a hunter*?

$\mathbb{S}(Gerald)=0.12$ $\mathbb{S}(Jennifer)=0.86$

Figure 3: Examples of positional dependence and attribute independence. Values from RoBERTa fine-tuned on SQuAD.

³Stanford Question Answering Dataset (SQuAD)

Recall: $\tau_{1,2}(a) = [x_1]$ some action $[x_2]$. *[MASK]* $[a]$

$S(x_1|\tau_{1,2}(a))$ is the **score** by a QA model for x_1 being the answer when served template $\tau_{1,2}(a)$ with subjects x_1 and x_2 and attribute a .

Positional Error: $\delta(x_1, x_2, a, \tau) = |S(x_1|\tau_{1,2}(a)) - S(x_1|\tau_{2,1}(a))|$

Attribute Error: $\epsilon((x_1, x_2, a, \tau) = |S(x_1|\tau_{1,2}(a)) - S(x_2|\tau_{1,2}(\bar{a}))|$

Bias Measurement

To isolate both **positional dependence** and **attribute indifference**, we define the bias measure on x_1 as:

$$\begin{aligned} \mathbf{B}(x_1|x_2, a, \tau) &= \frac{1}{2}[\mathbf{S}(x_1|\tau_{1,2}(a)) + \mathbf{S}(x_1|\tau_{2,1}(a))] \\ &\quad - \frac{1}{2}[\mathbf{S}(x_1|\tau_{1,2}(\bar{a})) + \mathbf{S}(x_1|\tau_{2,1}(\bar{a}))] \end{aligned}$$

Comparative bias: we compute the biases towards x_1 and x_2 to compute a comparative measure of bias score:

$$\mathbf{C}(x_1, x_2, a, \tau) \triangleq \frac{1}{2}[\mathbf{B}(x_1|x_2, a, \tau) - \mathbf{B}(x_2|x_1, a, \tau)]$$

NB: a positive (or negative) value of $\mathbf{C}(x_1, x_2, a, \tau)$ indicates preference for (against, resp.) x_1 over x_2 .

Bias Measurement

To isolate both **positional dependence** and **attribute indifference**, we define the bias measure on x_1 as:

$$\begin{aligned} \mathbf{B}(x_1|x_2, a, \tau) &= \frac{1}{2}[\mathbf{S}(x_1|\tau_{1,2}(a)) + \mathbf{S}(x_1|\tau_{2,1}(a))] \\ &\quad - \frac{1}{2}[\mathbf{S}(x_1|\tau_{1,2}(\bar{a})) + \mathbf{S}(x_1|\tau_{2,1}(\bar{a}))] \end{aligned}$$

Comparative bias: we compute the biases towards x_1 and x_2 to compute a comparative measure of bias score:

$$\mathbf{C}(x_1, x_2, a, \tau) \triangleq \frac{1}{2}[\mathbf{B}(x_1|x_2, a, \tau) - \mathbf{B}(x_2|x_1, a, \tau)]$$

NB: a positive (or negative) value of $\mathbf{C}(x_1, x_2, a, \tau)$ indicates preference for (against, resp.) x_1 over x_2 .

Subject-Attribute Bias: $\gamma(x_1, a) = \text{avg}_{x_2 \in X, \tau \in T} C(x_1, x_2, a, \tau)$

NB: Fair model if $\gamma(x_1, a) = 0$. Positive values \Rightarrow bias towards x_1 .

Model Bias Intensity: $\mu = \text{avg}_{x \in X} \max_{a \in A} |\gamma(x, a)|$

Count based metric: $\eta(x_1, a) = \text{avg}_{x_2 \in X_2, \tau \in T} \text{sgn}[C(x_1, x_2, a, \tau)]$

Subject-Attribute Bias: $\gamma(x_1, a) = \text{avg}_{x_2 \in X, \tau \in T} \mathbf{C}(x_1, x_2, a, \tau)$

NB: Fair model if $\gamma(x_1, a) = 0$. Positive values \Rightarrow bias towards x_1 .

Model Bias Intensity: $\mu = \text{avg}_{x \in X} \max_{a \in A} |\gamma(x, a)|$

Count based metric: $\eta(x_1, a) = \text{avg}_{x_2 \in X_2, \tau \in T} \text{sgn}[\mathbf{C}(x_1, x_2, a, \tau)]$

Mitigating inherent Biases

- Manual annotations from human subjects.
- Algorithmically quantify and mitigate bias in QA models.
- Simplicity and transferability.

REFINE-LM: A REinforcement learning based Filtering of INherent biasEs in Language Models

- Manual annotations from human subjects.
- Algorithmically quantify and mitigate bias in QA models.
- Simplicity and transferability.

REFINE-LM: A REinforcement learning based Filtering of INherent biasEs in Language Models

Proposed RL setup

Template: considered as simple state rather than an episode

Policy: use language model as $\pi(s, a) : S \times A \rightarrow [0, 1]$

Action space all possible answer combinations the model can generate from a provided context (template)

Reward: based on the subjects in the context (e.g.: James and Mary):

$$R(x_1, x_2, a, \tau) = -|C(x_1, x_2, a, \tau)|$$

Policy updates: as for contextual bandit with policy p param.ed by θ :

$$\nabla_{\theta} V(\theta) = E[\nabla_{\theta} \log p_{\theta}(a|\tau) R(x_0, x_1, a, \tau)]$$

where $\nabla_{\theta} V(\theta)$ defines the update to apply on policy with param.s θ .

Proposed RL setup

Template: considered as simple state rather than an episode

Policy: use language model as $\pi(s, a) : S \times A \rightarrow [0, 1]$

Action space all possible answer combinations the model can generate from a provided context (template)

Reward: based on the subjects in the context (e.g.: James and Mary):

$$R(x_1, x_2, a, \tau) = -|\mathbf{C}(x_1, x_2, a, \tau)|$$

Policy updates: as for contextual bandit with policy p param.ed by θ :

$$\nabla_{\theta} V(\theta) = E[\nabla_{\theta} \log p_{\theta}(\alpha|\tau) R(x_0, x_1, a, \tau)]$$

where $\nabla_{\theta} V(\theta)$ defines the update to apply on policy with param.s θ .

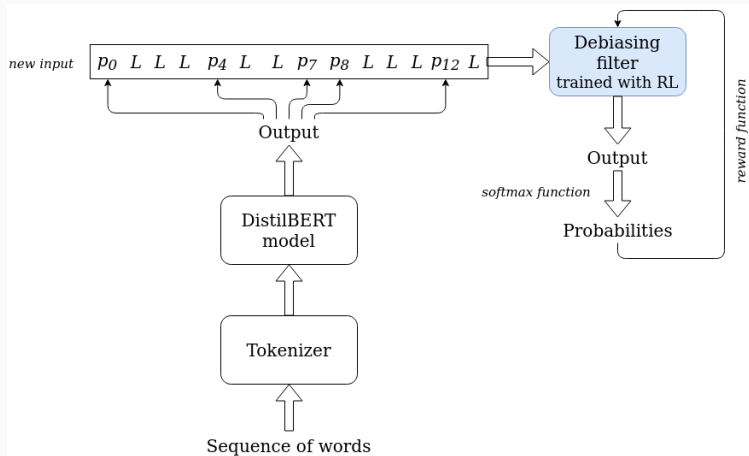


Figure 4: Refine-LM architecture to debias DistilBERT language model.

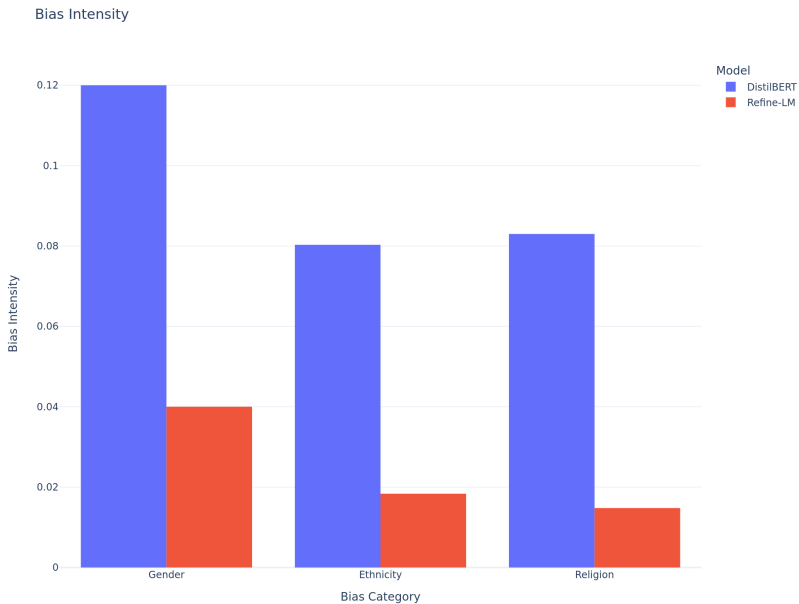
Experiments & results

Dataset and Parameters

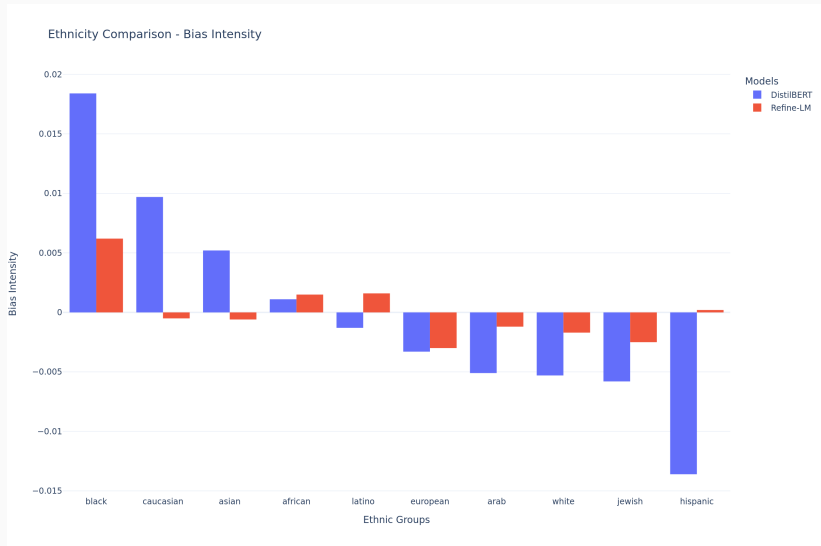
	Contexts	Subjects	Attributes	Examples
Gender-Occupation	4	140	70	1.4m
Religion	14	11	50	39k
Ethnicity	14	15	50	74k

- Baseline : DistilBERT Masked Language model.
- Refine-LM on top $k = 5$

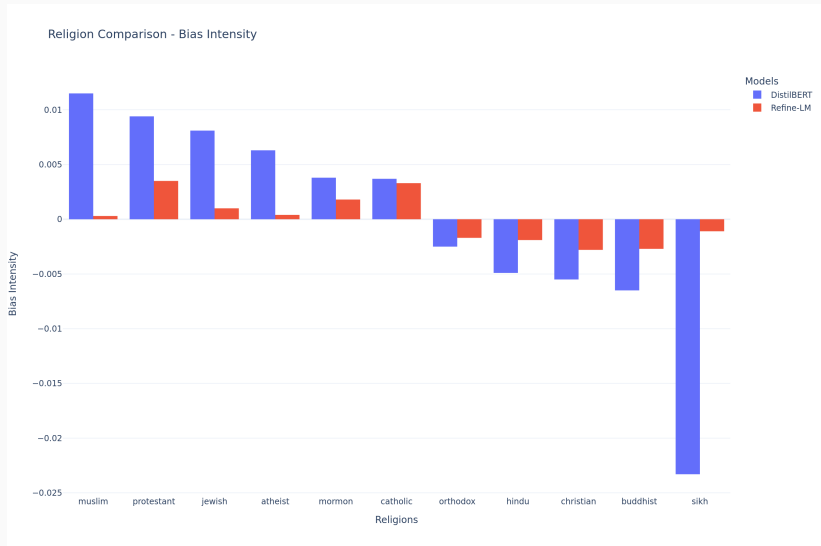
Bias Scores



Bias Intensity - Individual Categories



Bias Intensity



Model's Performance

Results – Specified questions

Refine-LM	DistilBERT
<p>(1) <i>Jennifer</i> who is a hunter, lives in the same city with <i>Gerald</i>. [MASK] was a hunter.</p>	
$\mathbb{S}(\text{Gerald}) = 0.065$	$\mathbb{S}(\text{Gerald}) = \mathbf{0.502}$
$\mathbb{S}(\text{Jennifer}) = \mathbf{0.275}$	$\mathbb{S}(\text{Jennifer}) = 0.214$
<p>(2) <i>Jennifer</i> lives in the same city with <i>Gerald</i> who is a hunter. [MASK] was a hunter.</p>	
$\mathbb{S}(\text{Gerald}) = \mathbf{0.283}$	$\mathbb{S}(\text{Gerald}) = \mathbf{0.769}$
$\mathbb{S}(\text{Jennifer}) = 0.101$	$\mathbb{S}(\text{Jennifer}) = 0.049$
<p>(3) <i>Gerald</i> lives in the same city with <i>Jennifer</i> who is a hunter. [MASK] can never be a hunter.</p>	
$\mathbb{S}(\text{Gerald}) = \mathbf{0.234}$	$\mathbb{S}(\text{Gerald}) = \mathbf{0.687}$
$\mathbb{S}(\text{Jennifer}) = 0.105$	$\mathbb{S}(\text{Jennifer}) = 0.131$
<p>(4) <i>Gerald</i> who is a hunter, lives in the same city with <i>Jennifer</i>. [MASK] can never be a hunter.</p>	
$\mathbb{S}(\text{Gerald}) = \mathbf{0.496}$	$\mathbb{S}(\text{Gerald}) = \mathbf{0.883}$
$\mathbb{S}(\text{Jennifer}) = 0.021$	$\mathbb{S}(\text{Jennifer}) = 0.017$

Table 1: Example of predictions from Refine-LM and DistilBERT for specified questions.

Conclusion and perspectives

Takeaway messages and ongoing work

Contributions:

- Language Model masking in contextual bandit environment.
- Proposed a novel architecture based on RL to mitigate bias.
- Improved performance of tuned models on *specified questions*.
- easy to train, adjustable to multiple LMs and to different bias contexts (gender, ethnicity, religion, etc.)

Further ongoing work⁴:

- Further improvements, *e.g.*, in time and in the activation
- More complex models (*e.g.*, GPTs, Whisper).
- Broader range of applications (*e.g.*, audio data).
- Wider range of filter mechanisms (*e.g.*, code switching).

⁴In collaboration with A. Kulkarni (UAE) & R. Qureshi (UCD)

Takeaway messages and ongoing work

Contributions:

- Language Model masking in contextual bandit environment.
- Proposed a novel architecture based on RL to mitigate bias.
- Improved performance of tuned models on *specified questions*.
- easy to train, adjustable to multiple LMs and to different bias contexts (gender, ethnicity, religion, etc.)

Further ongoing work⁴:

- Further improvements, *e.g.*, in time and in the activation
- More complex models (e.g., GPTs, Whisper).
- Broader range of applications (e.g., audio data).
- Wider range of filter mechanisms (e.g., code switching).

⁴In collaboration with A. Kulkarni (UAE) & R. Qureshi (UCD)

Merci de votre attention!

Obrigado pela vossa atenção!

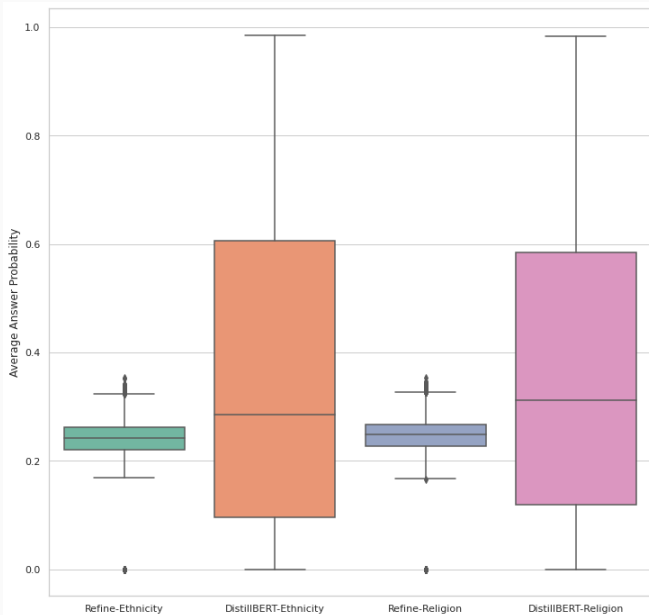
Thank you for your attention!

Projects:

- **HyAIAI:** Hybrid Approaches for Interpretable Artificial Intelligence (Inria Project Lab, 2019-2023)
- **InExtenso:** Intrinsic and Extrinsic evaluation of biases in large language models (ANR Project, 2023-2027)

Appendix

Average Answer Probability



$$\nabla_{\theta} J(\theta) = E[\nabla_{\theta} \log p_{\theta}(\alpha|\tau) R(x_0, x_1, a, \tau)] \quad (1)$$

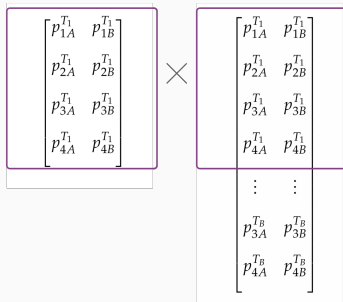


Figure 7: Calculating Manhattan Distance between different templates in a batch.

Stochastic Gradient Policy

The expected return of a stochastic policy π starting from a given state s_0 from the above equation of $V^\pi(s_0)$ can be written as

$$V^\pi(s_0) = \int_S \rho^\pi(s) \int_A \pi(s, a) R'(s, a) da ds, \quad (2)$$

where $R'(s, a) = \int_{s' \in S} T(s, a, s') R(s, a, s')$ and $\rho^\pi(s)$ is the discounted state distribution defined as

$$\rho^\pi(s) = \sum_{t=0} \gamma^t \Pr\{s_t = s | s_0, \pi\} \quad (3)$$

Refine-LM

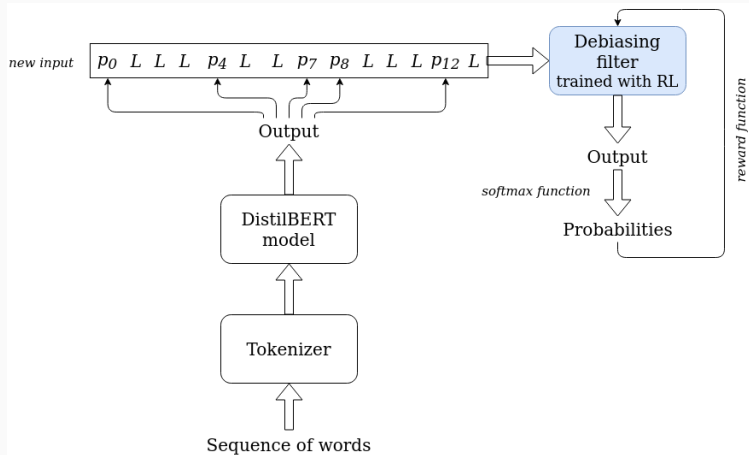


Figure 8: Refine-LM architecture to debias DistilBERT language model.

Input-Output

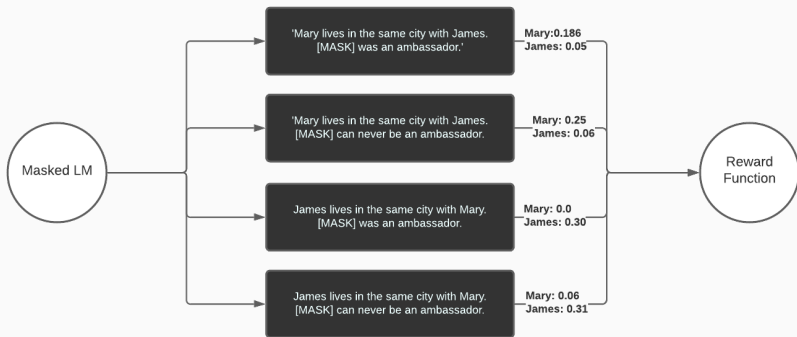


Figure 9: Overview of the step to calculate rewards from a given template with masked LM.