Unsupervised Editing for Counterfactual Stories

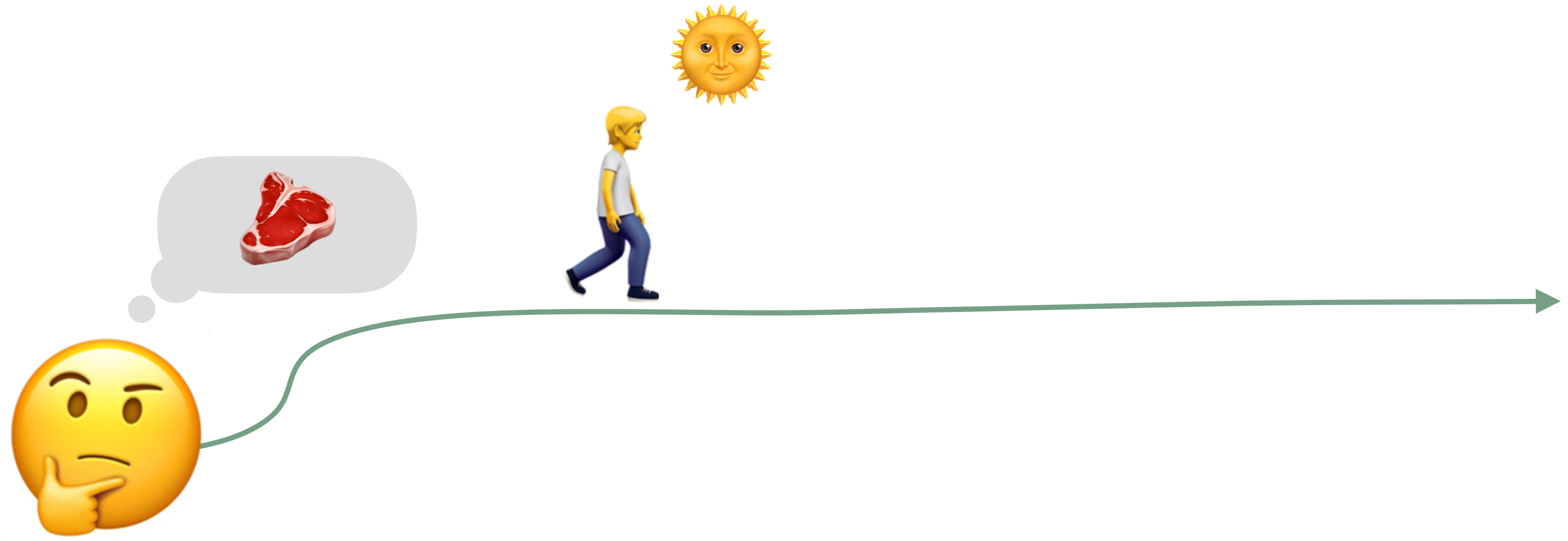
Jiangjie Chen[1,3], Chun Gan[2], Sijie Cheng[1], Hao Zhou[3], Yanghua Xiao[1], Lei Li[4]

复旦大学 FUDAN UNIVERSITY
JD.COM
ByteDance
UC SANTA BARBARA

Automatic Story Writing

“I want some steak🤤!”

“It’s a sunny day, let’s go out😃!”



"Nice steak they have😋!"

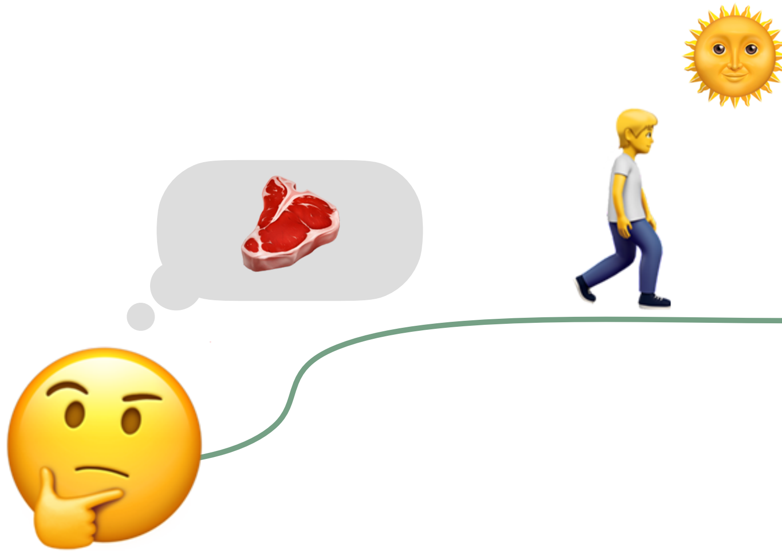


Photo taken @Shanghai

Automatic Story *Re*-Writing

Photo taken @Shanghai

"What if ?"

“Oh☹️, I hate rainy days.”

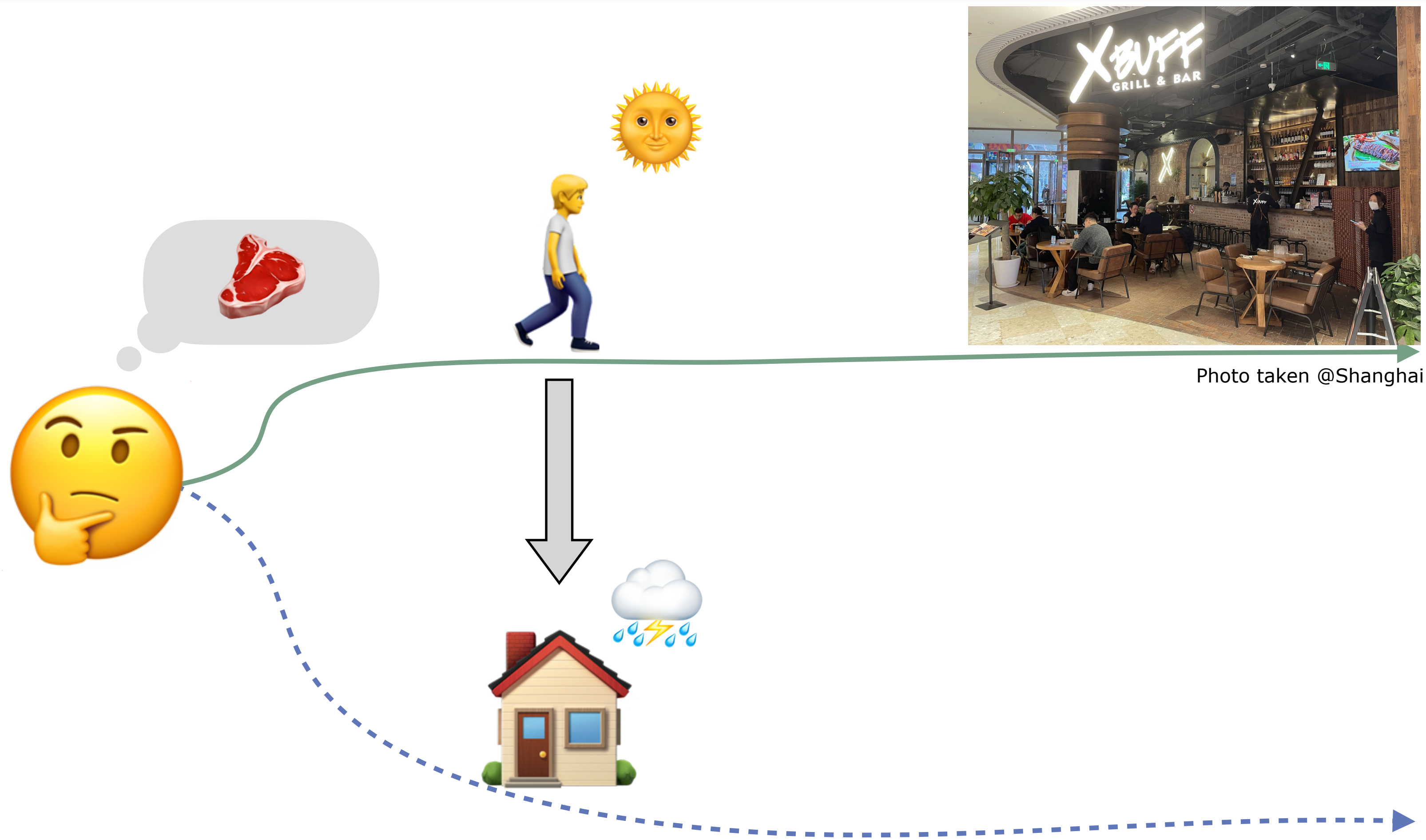


Photo taken @Shanghai

What if…?

“What should I do🤨?”

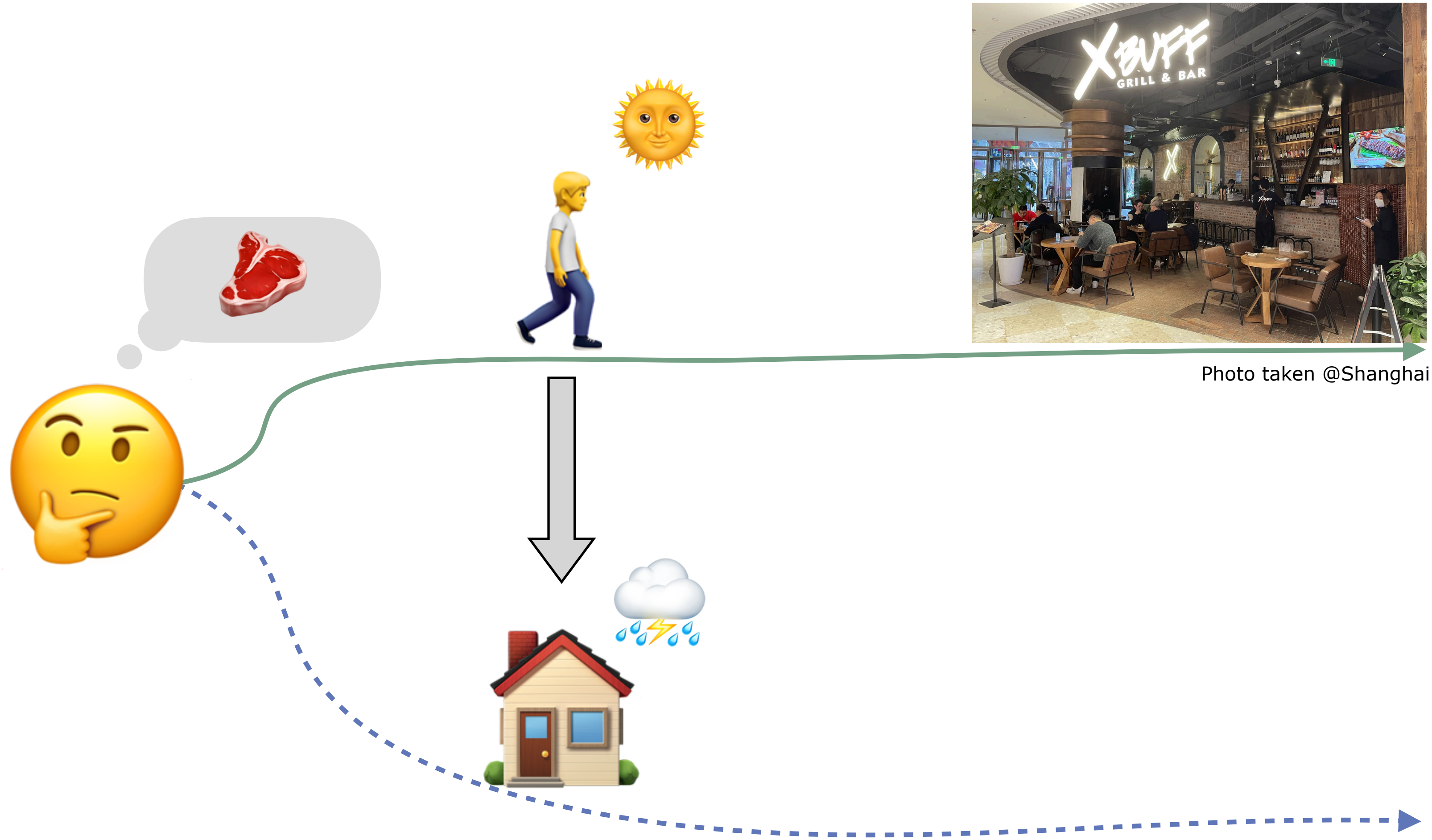


Photo taken @Shanghai

What if…?

“I might as well cook it myself!”

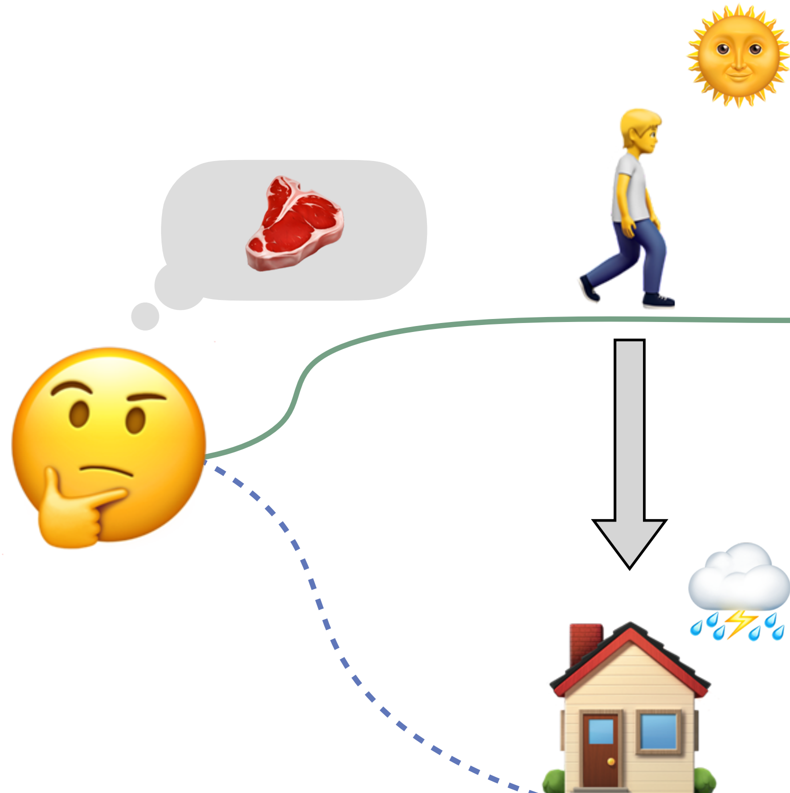


Photo taken @Shanghai

Photo credit to Dr. Yemeng ZHOU

What if...?

Counterfactual Story Rewriting for Creative NLG

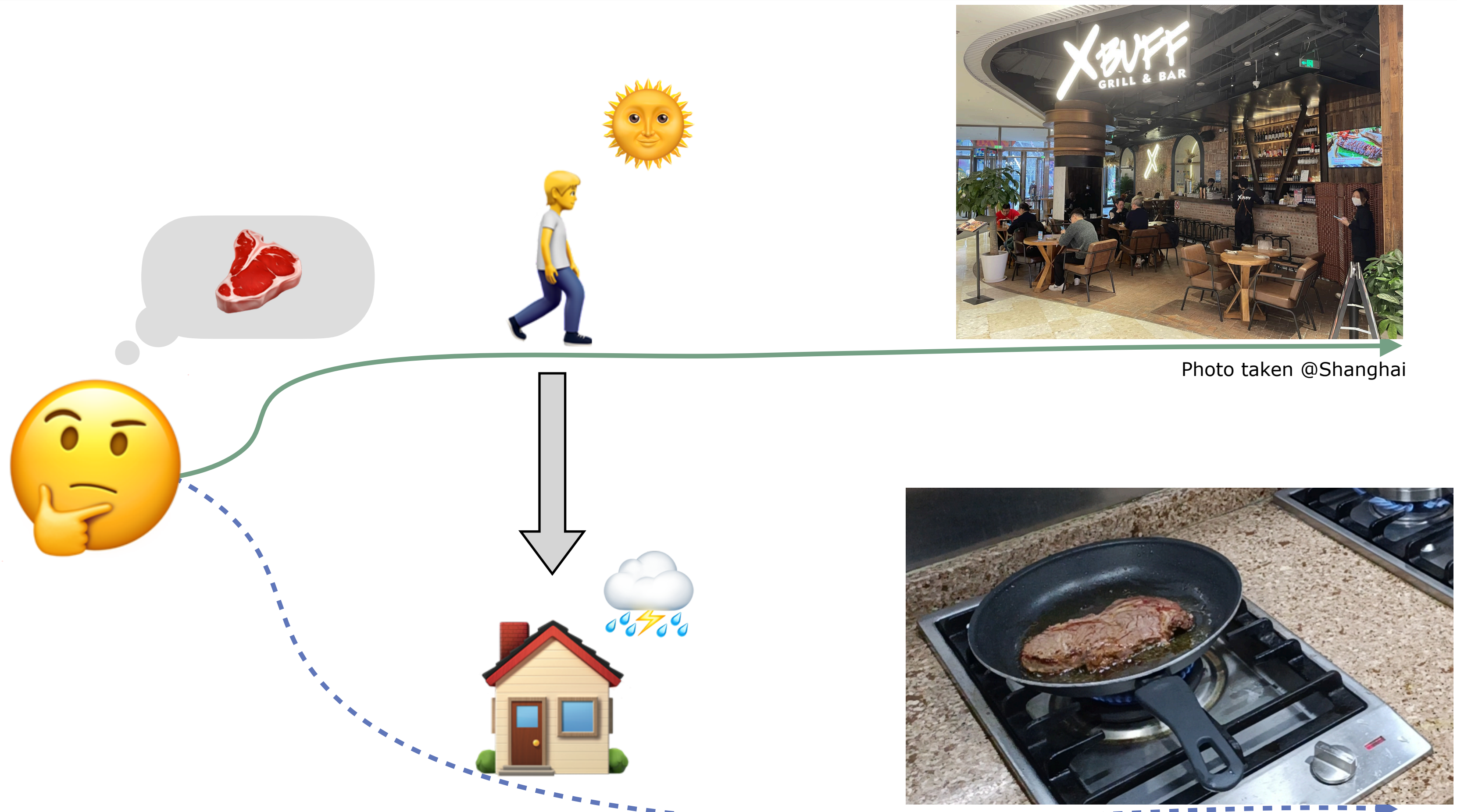


Photo taken @Shanghai

Photo credit to Dr. Yemeng ZHOU

What if...?

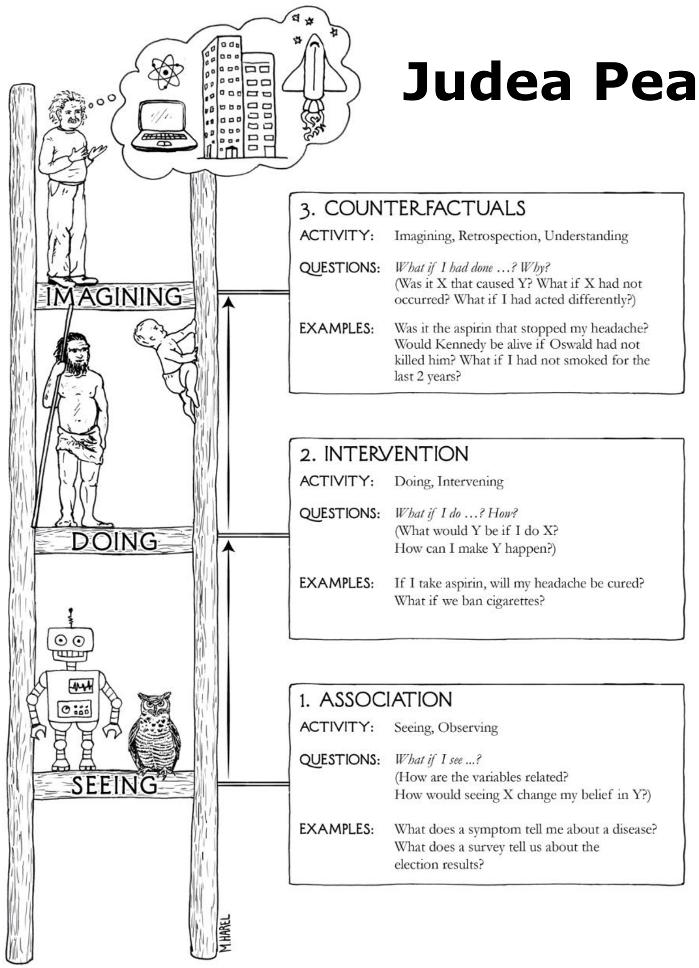
Counterfactual Reasoning

- A hypothetical thinking process to assess possible outcomes by modifying certain prior conditions.

Counterfactual Reasoning

- A hypothetical thinking process to assess possible outcomes by modifying certain prior conditions.

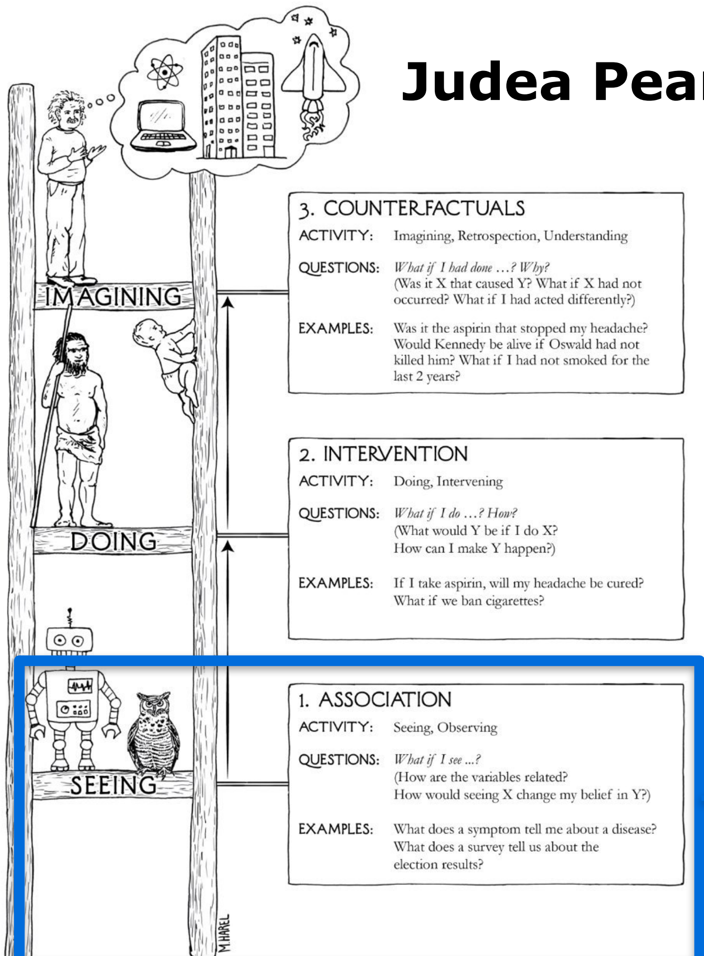
Judea Pearl's "Ladder of Causality"



Counterfactual Reasoning

- A hypothetical thinking process to assess possible outcomes by modifying certain prior conditions.

Judea Pearl's "Ladder of Causality"

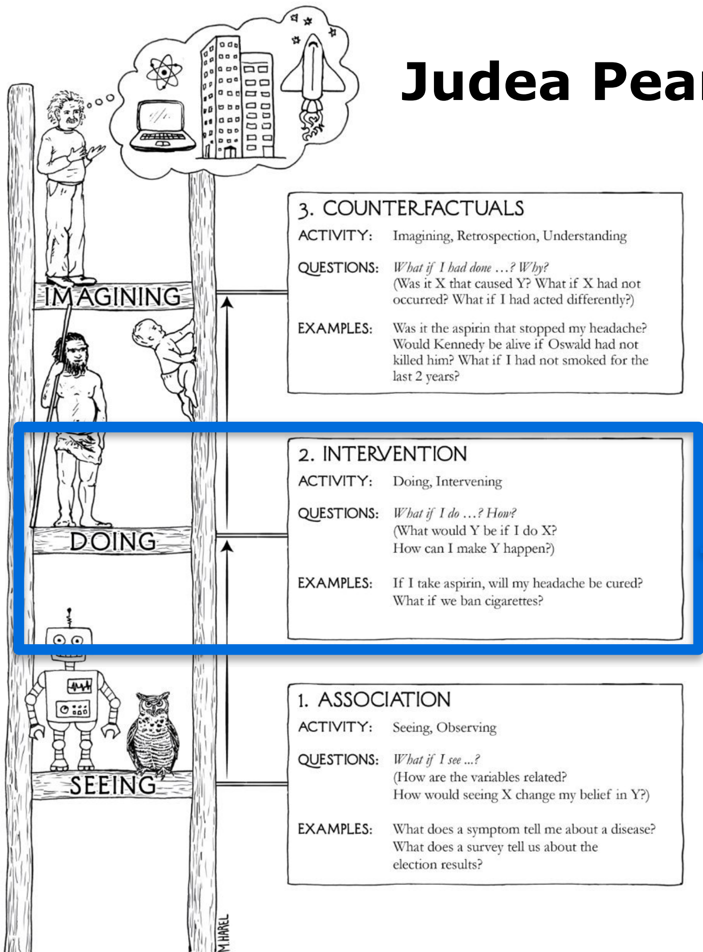


Association: What if I see...?

Counterfactual Reasoning

- A hypothetical thinking process to assess possible outcomes by modifying certain prior conditions

Judea Pearl's "Ladder of Causality"

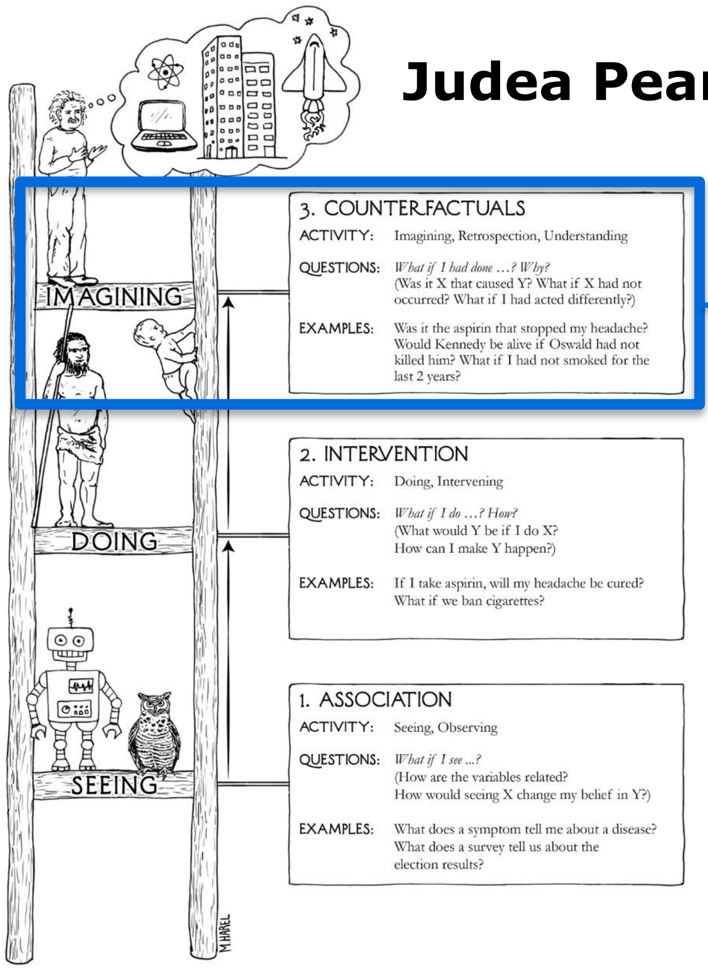


Intervention: What if I do...?

Counterfactual Reasoning

- A hypothetical thinking process to assess possible outcomes by modifying certain prior conditions

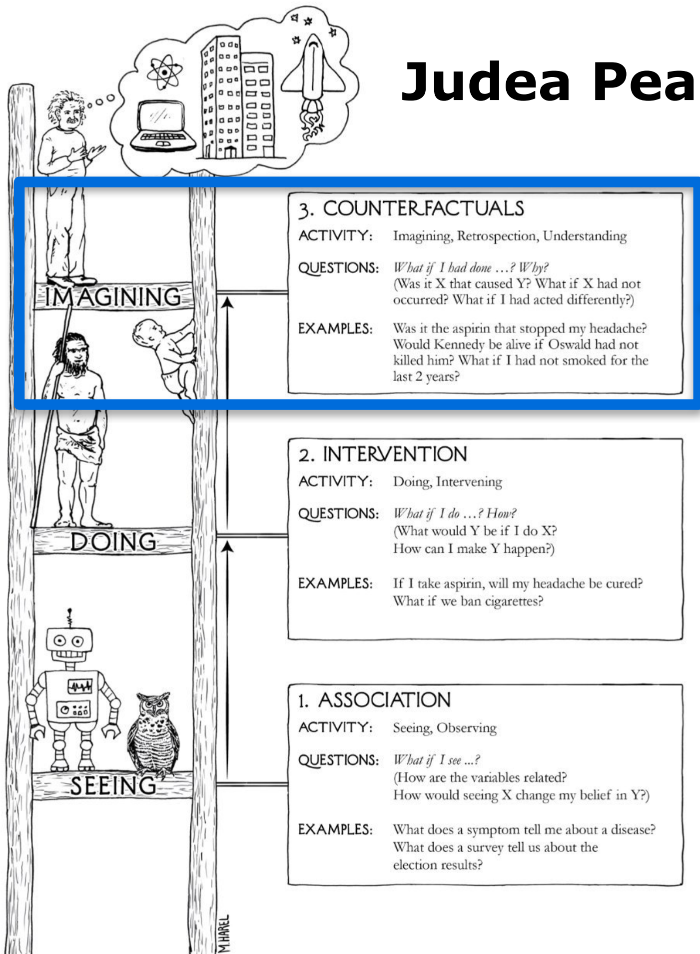
Judea Pearl's "Ladder of Causality"



Counterfactuals: What if I had done…?

Counterfactual Reasoning

- A hypothetical thinking process to assess possible outcomes by modifying certain prior conditions



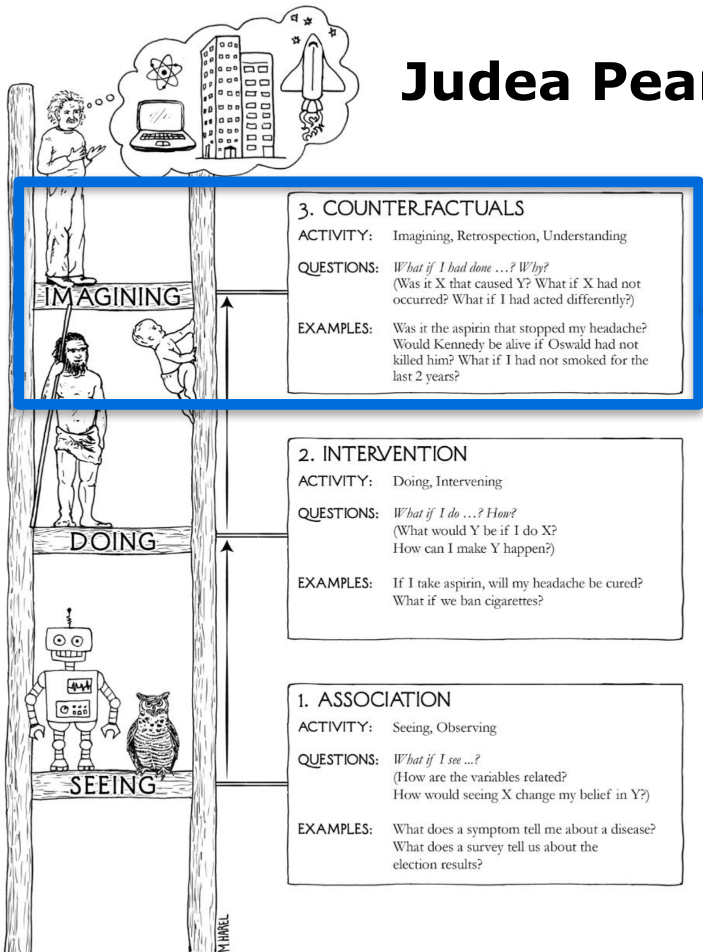
Judea Pearl's "Ladder of Causality"

Counterfactuals: What if I had done…?

- ***Challenge:*** *Causal Invariance*
  - the factors that hold constant with the change of conditions in a series of events

Counterfactual Reasoning

- A hypothetical thinking process to assess possible outcomes by modifying certain prior conditions



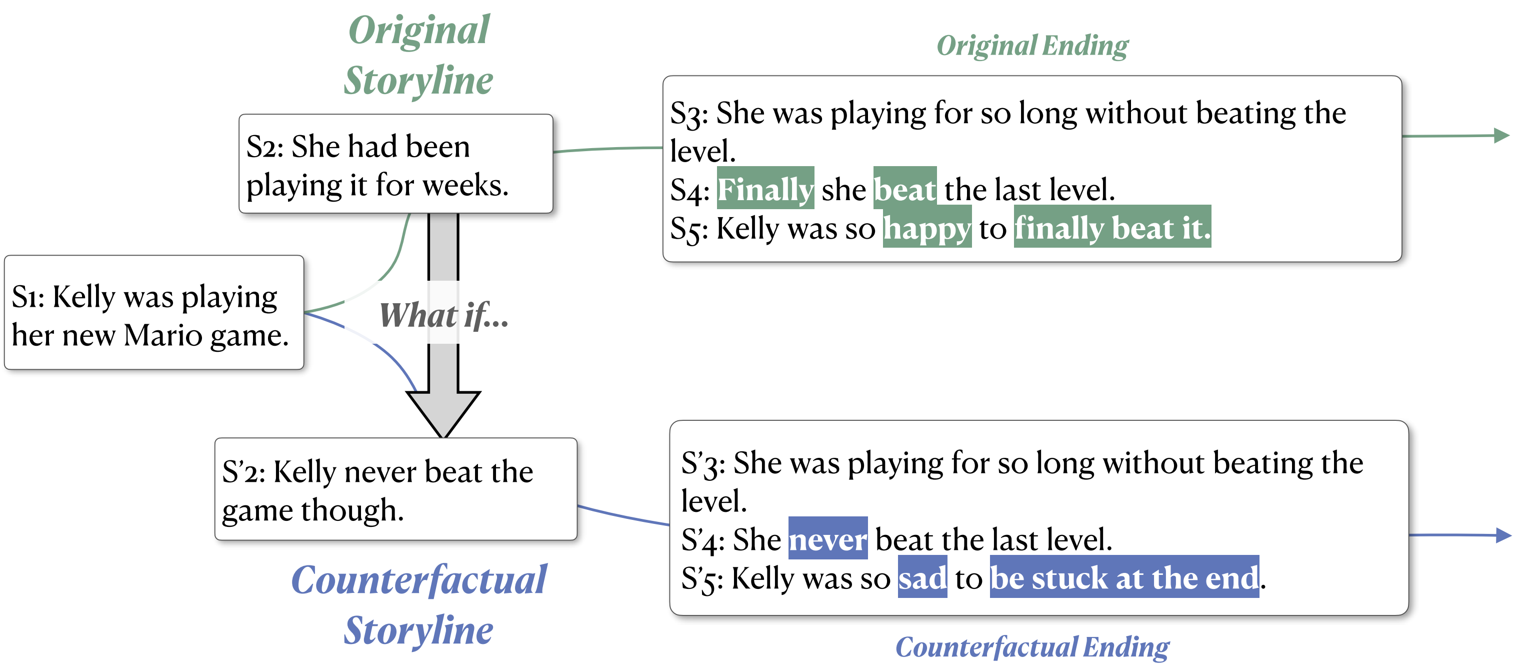
Judea Pearl's "Ladder of Causality"

Counterfactuals: What if I had done...?

- ***Challenge:*** *Causal Invariance*
  - the factors that hold constant with the change of conditions in a series of events

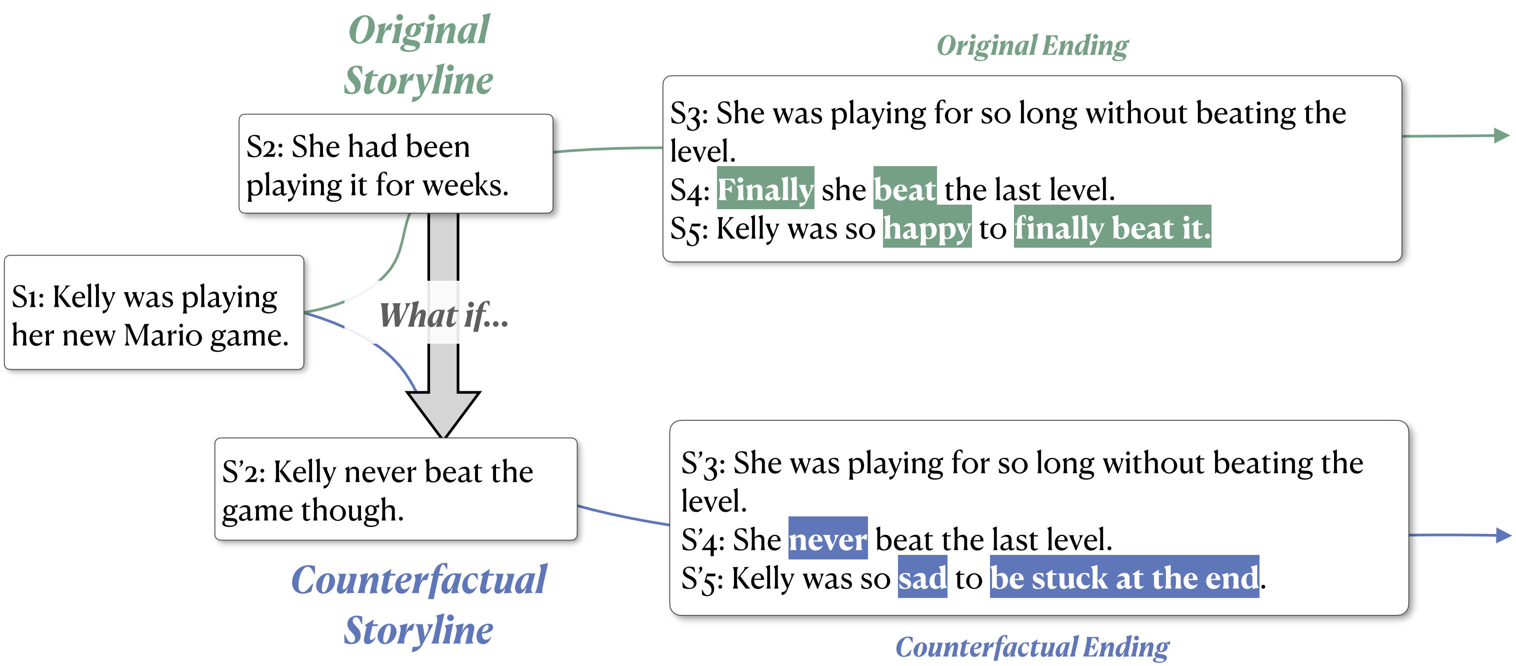
The *Trade-off*: Minimal-edits vs. Coherence

*Can we rewrite a new story ending with **minimal edits**?*



The *Trade-off*: Minimal-edits vs. Coherence

Can we rewrite a new story ending with ***minimal edits****?*



For ***pre-trained LMs****, massive editing can almost certainly lead to a coherent ending.*

The *Trade-off*: Minimal-edits vs. Coherence

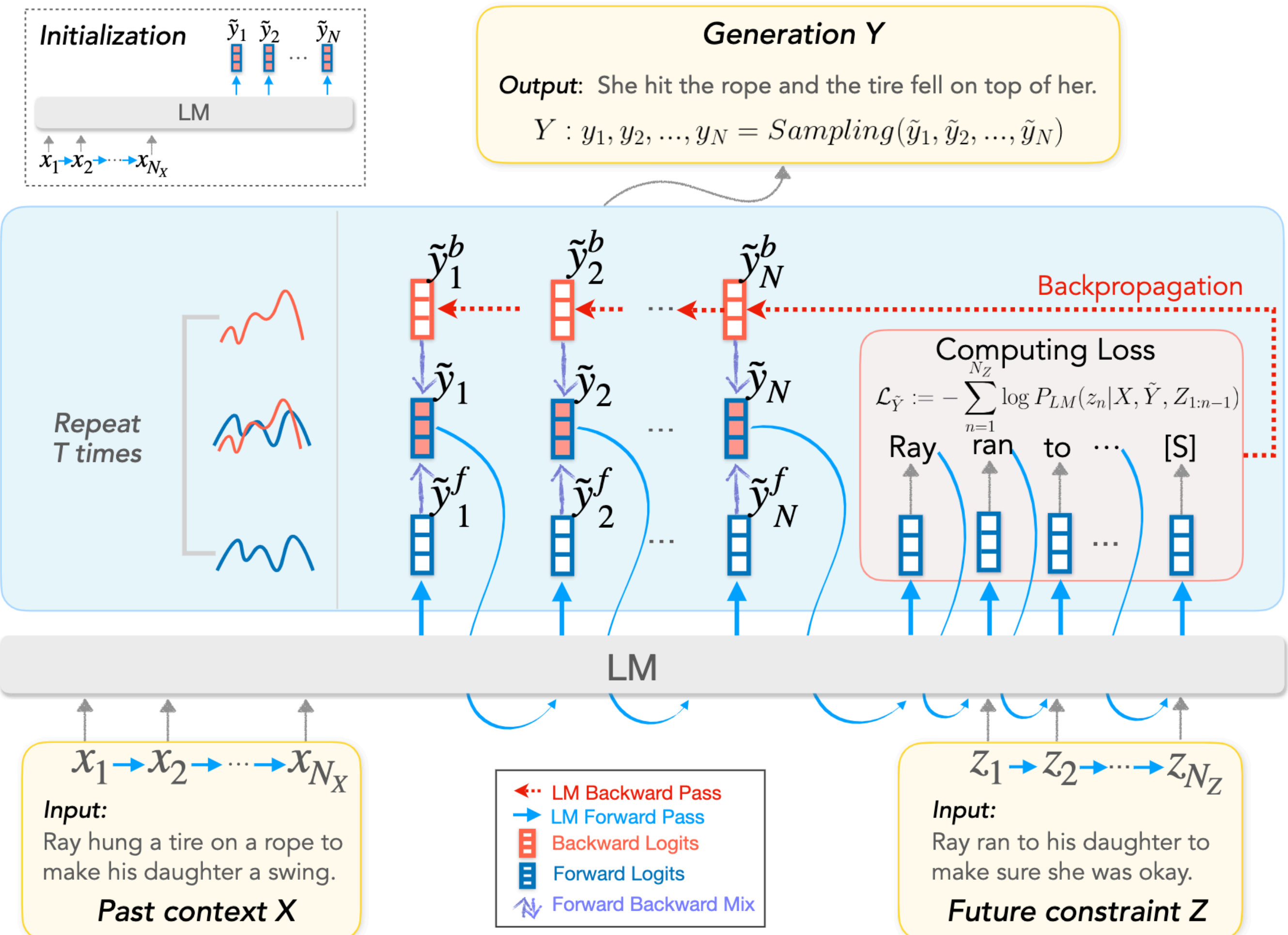
*Can we rewrite a new story ending with **minimal edits**?*

Also do it without supervision!

Humans do not need training to imagine possible futures!

*For **pre-trained LMs**, massive editing can almost certainly lead to a coherent ending.*

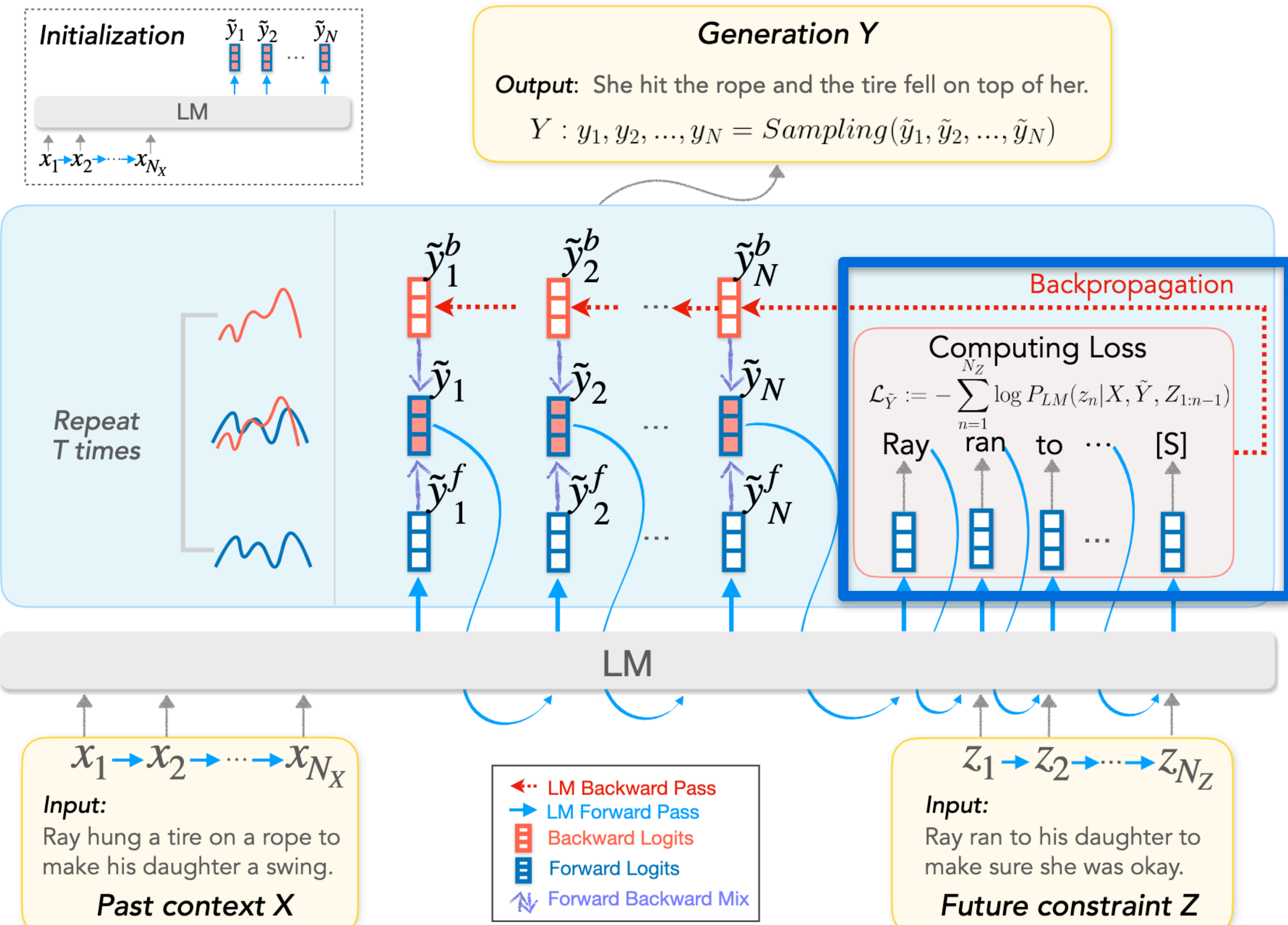
How does Previous Method Solve this Problem?



DELOREAN

Qin, Lianhui, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. **Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning.** EMNLP 2020

How does Previous Method Solve this Problem?

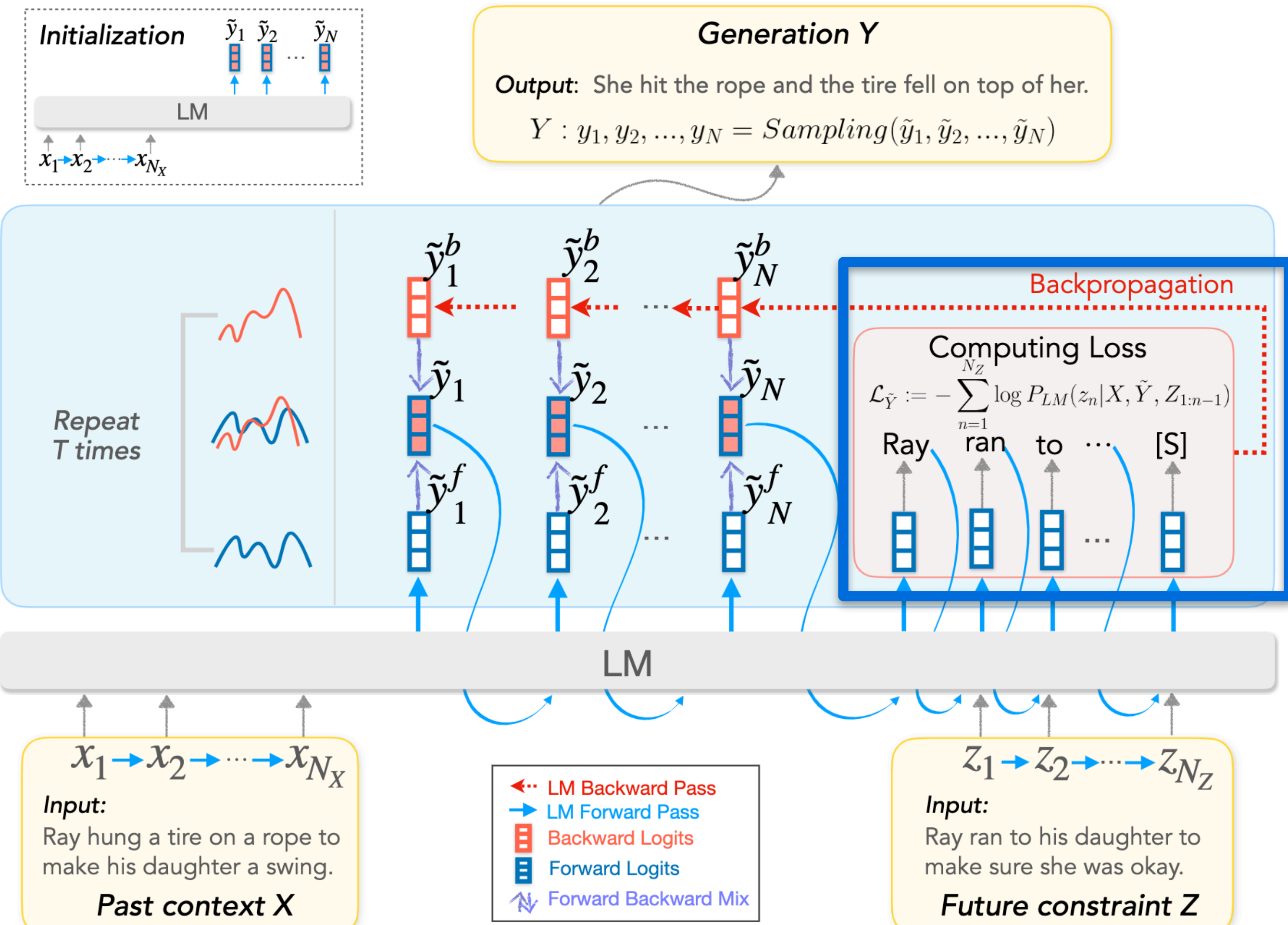


Constraints backprop to the pre-trained LMs

DELOREAN

Qin, Lianhui, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. **Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning.** EMNLP 2020

How does Previous Method Solve this Problem?



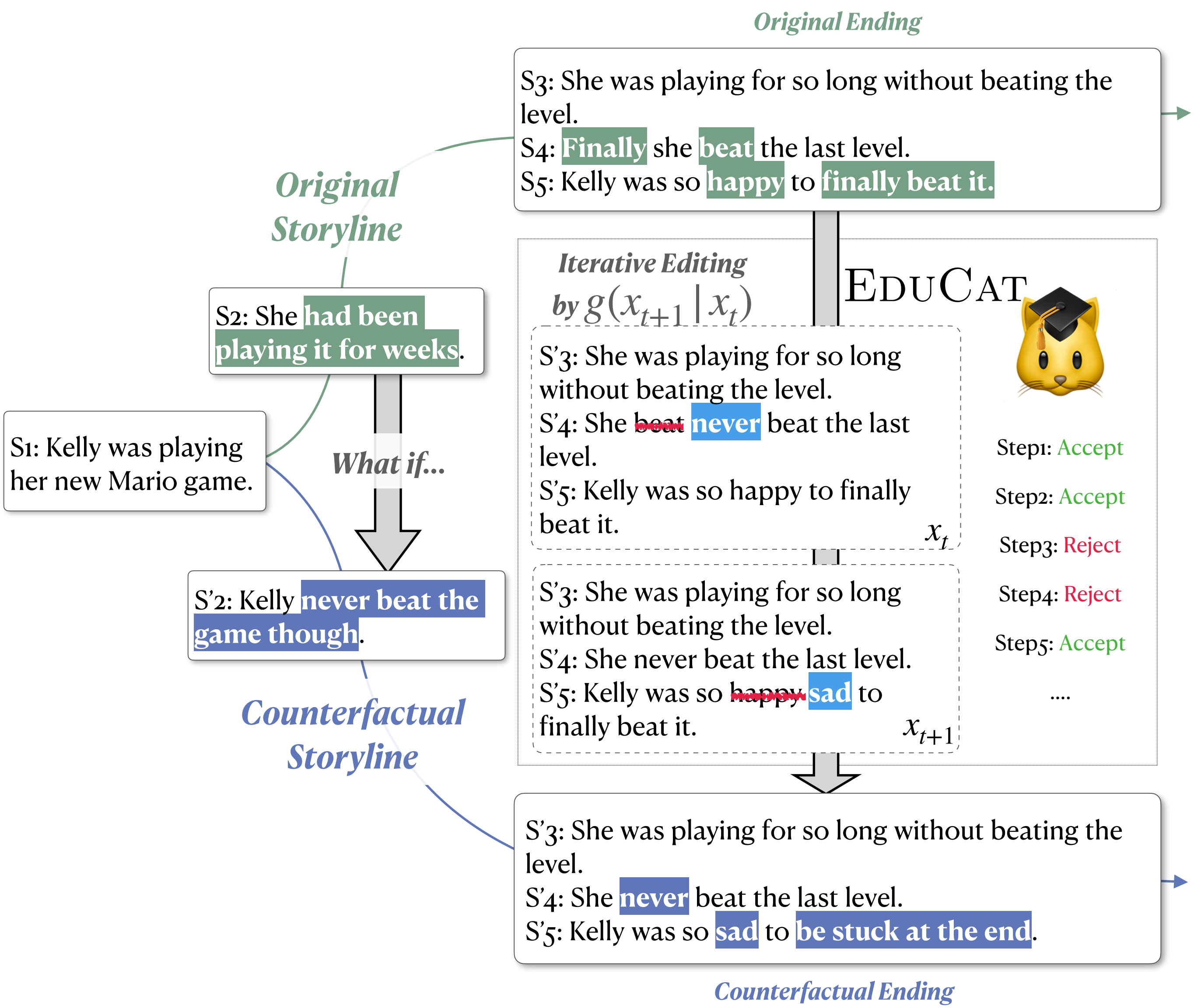
Constraints backprop to the pre-trained LMs

Still massive edits?

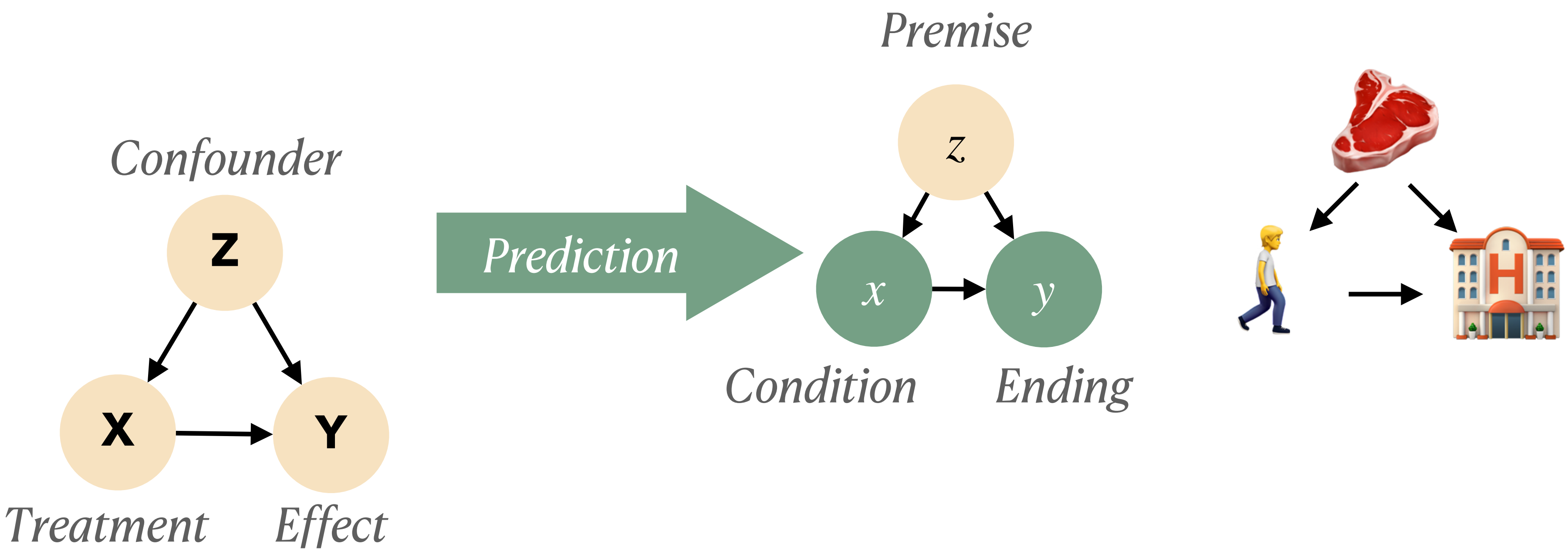
DELOREAN

Qin, Lianhui, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. **Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning.** EMNLP 2020

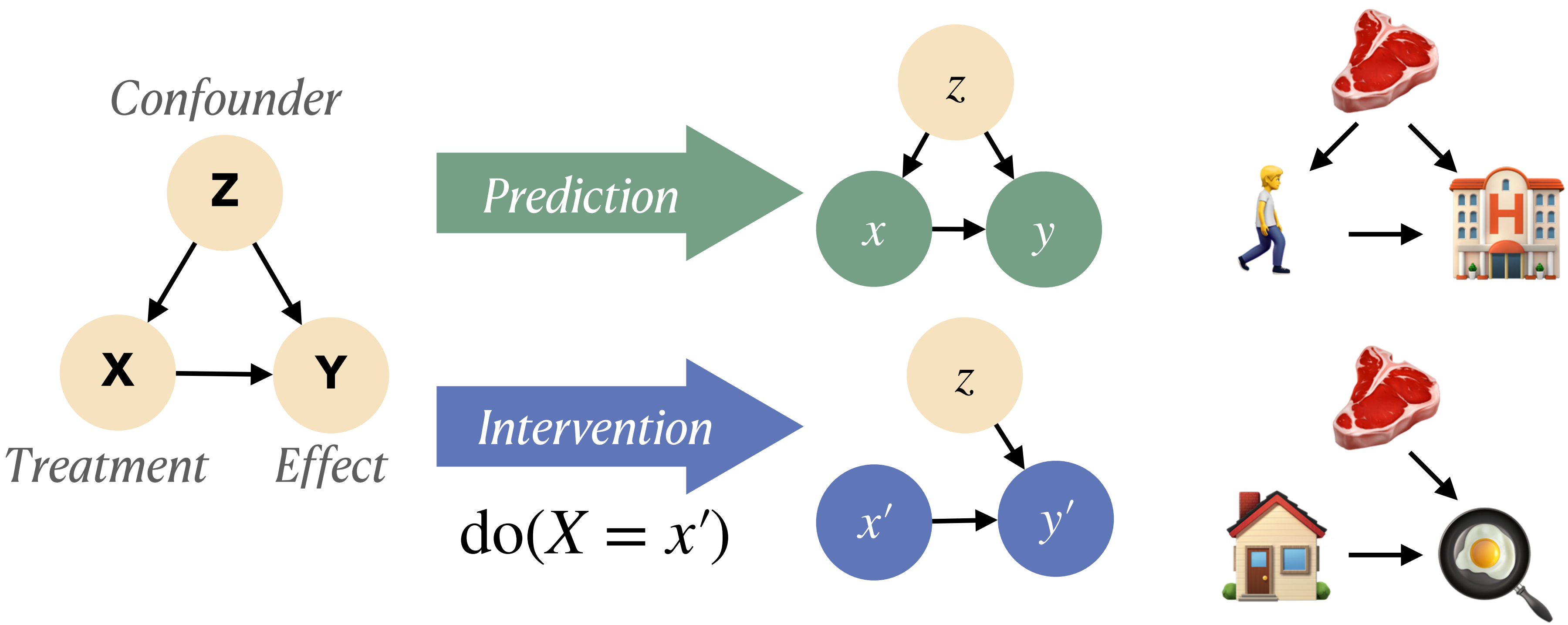
EDUCAT: Edit a Story Ending



Structured Causal Model



Structured Causal Model



Estimating Potential Outcome After Intervention — Causal Risk Ratio

Causal Risk Ratio:

$$\mathrm{CRR} = \frac{\mathrm{P}(Y = y \mid \mathrm{do}(X = x'), Z = z)}{\mathrm{P}(Y = y \mid \mathrm{do}(X = x), Z = z)}$$

$$P(Y = y \mid \mathrm{do}(X = x')) = \sum_{z} P(Y = y \mid X = x', Z = z) P(Z = z)$$

Estimating Potential Outcome After Intervention — Causal Risk Ratio

Causal Risk Ratio:

$$\text{CRR} = \frac{\mathrm{P}(Y = y \mid \mathrm{do}(X = x'), Z = z)}{\mathrm{P}(Y = y \mid \mathrm{do}(X = x), Z = z)}$$

$$P(Y = y \mid \mathrm{do}(X = x')) = \sum_{z} P(Y = y \mid X = x', Z = z) P(Z = z)$$

Causal Sufficiency Assumption

$$P(Y = y \mid \mathrm{do}(X = x)) = P(Y = y \mid X = x, Z = z)$$

Estimating Potential Outcome After Intervention — Causal Risk Ratio

Causal Risk Ratio:

$$\mathrm{CRR} = \frac{\mathrm{P}(Y = y \mid \mathrm{do}(X = x'), Z = z)}{\mathrm{P}(Y = y \mid \mathrm{do}(X = x), Z = z)}$$

$$P(Y = y \mid \mathrm{do}(X = x')) = \sum_{z} P(Y = y \mid X = x', Z = z) P(Z = z)$$

Causal Sufficiency Assumption

$$P(Y = y \mid \mathrm{do}(X = x)) = P(Y = y \mid X = x, Z = z)$$

$$\mathrm{CRR} = \frac{\mathrm{P}(Y = y \mid X = x', Z = z)}{\mathrm{P}(Y = y \mid X = x, Z = z)}$$ 🙋‍♂️

Unsupervised Constrained Editing via MCMC Sampling

- CGMH: sentence generation with **Metropolis-Hastings Sampling**. [Miao et al. 2019]

Miao, Ning, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. **CGMH: Constrained sentence generation by Metropolis-Hastings sampling**. AAAI 2019.

Unsupervised Constrained Editing via MCMC Sampling

- CGMH: sentence generation with **Metropolis-Hastings Sampling**. [Miao et al. 2019]
  - Define desired properties as stationary distribution $\pi(\mathrm{y})$

Unsupervised Constrained Editing via MCMC Sampling

- CGMH: sentence generation with **Metropolis-Hastings Sampling**. [Miao et al. 2019]
  - Define desired properties as stationary distribution $\pi(\mathrm{y})$
  - Move y_t to y_{t+1} by generating from the proposal distribution $g(\mathrm{y}_{t+1} \mid \mathrm{y}_t)$

Unsupervised Constrained Editing via MCMC Sampling

- CGMH: sentence generation with **Metropolis-Hastings Sampling**. [Miao et al. 2019]
  - Define desired properties as stationary distribution $\pi(\mathrm{y})$
  - Move y_t to y_{t+1} by generating from the proposal distribution $g(\mathrm{y}_{t+1} | \mathrm{y}_t)$
  - Accept a proposal with acceptance rate $\alpha(\mathrm{y}_{t+1} | \mathrm{y}_t)$

Unsupervised Constrained Editing via MCMC Sampling

- CGMH: sentence generation with **Metropolis-Hastings Sampling**. [Miao et al. 2019]
  - Define desired properties as stationary distribution $\pi(\mathrm{y})$
  - Move y_t to y_{t+1} by generating from the proposal distribution $g(\mathrm{y}_{t+1} | \mathrm{y}_t)$
  - Accept a proposal with acceptance rate $\alpha(\mathrm{y}_{t+1} | \mathrm{y}_t)$
  - Iterate until convergence

Unsupervised Constrained Editing via MCMC Sampling

- CGMH: sentence generation with **Metropolis-Hastings Sampling**. [Miao et al. 2019]
  - Define desired properties as stationary distribution $\pi(\mathrm{y})$
  - Move y_t to y_{t+1} by generating from the proposal distribution $g(\mathrm{y}_{t+1} | \mathrm{y}_t)$
  - Accept a proposal with acceptance rate $\alpha(\mathrm{y}_{t+1} | \mathrm{y}_t)$
  - Iterate until convergence
  - Rank the accepted ones with $\pi(\cdot)$

Unsupervised Constrained Editing via MCMC Sampling

- CGMH: sentence generation with **Metropolis-Hastings Sampling**. [Miao et al. 2019]
  - Define desired properties as stationary distribution $\pi(\mathrm{y})$
  - Move y_t to y_{t+1} by generating from the proposal distribution $g(\mathrm{y}_{t+1} \mid \mathrm{y}_t)$
  - Accept a proposal with acceptance rate $\alpha(\mathrm{y}_{t+1} \mid \mathrm{y}_t)$
  - Iterate until convergence
  - Rank the accepted ones with $\pi(\cdot)$

$$\alpha(\mathrm{y}_{t+1} \mid \mathrm{y}_t) = \min \left\{ 1, \frac{\pi(\mathrm{y}_{t+1})^{1/T} g(\mathrm{y}_t \mid \mathrm{y}_{t+1})}{\pi(\mathrm{y}_t)^{1/T} g(\mathrm{y}_{t+1} \mid \mathrm{y}_t)} \right\}$$

Unsupervised Constrained Editing via MCMC Sampling

- CGMH: sentence generation with **Metropolis-Hastings Sampling**. [Miao et al. 2019]
  - Define desired properties as stationary distribution $\pi(\mathrm{y})$
  - Move y_t to y_{t+1} by generating from the proposal distribution $g(\mathrm{y}_{t+1} | \mathrm{y}_t)$
  - Accept a proposal with acceptance rate $\alpha(\mathrm{y}_{t+1} | \mathrm{y}_t)$
  - Iterate until convergence
  - Rank the accepted ones with $\pi(\cdot)$

$$\alpha(\mathrm{y}_{t+1} | \mathrm{y}_t) = \min \left\{ 1 \; \frac{\pi(\mathrm{y}_{t+1})^{1/T} g(\mathrm{y}_t | \mathrm{y}_{t+1})}{\boxed{\pi(\mathrm{y}_t)^{1/T}} g(\mathrm{y}_{t+1} | \mathrm{y}_t)} \right\}$$

$$\pi(\mathrm{y}) \propto \mathcal{X}_{LM}(\mathrm{y}) \cdot \mathcal{X}_{Coh}(\mathrm{y})$$

coherence & fluency

Desired Properties: Fluency and Coherence

- **Fluency Score**
  - Sentence probability from a PLM (e.g., GPT-2)

$$\mathcal{X}_{\mathrm{LM}}(y^*) = \prod_{i=1}^{N} P_{\mathrm{LM}}(y_i^* \mid z, x', y_{<i}^*)$$

Desired Properties: Fluency and Coherence

- **Fluency Score**
  - Sentence probability from a PLM (e.g., GPT-2)

$$\mathcal{X}_{\mathrm{LM}}(y^*) = \prod_{i=1}^{N} P_{\mathrm{LM}}(y_i^* \mid z, x', y_{<i}^*)$$

- **Coherence Score**
  - Punish proposed endings contradictory to the counterfactual conditions but consistent with the initial ones
  - Inspired by CRR
  - P_{Coh} could be changed from a PLM to more sophisticated ones

$$\mathcal{X}_{\mathrm{Coh}}(y^*) = \frac{P_{\mathrm{Coh}}(Y = y^* \mid z, x')}{P_{\mathrm{Coh}}(Y = y^* \mid z, x)}$$

$$\mathrm{CRR} = \frac{\mathrm{P}(Y = y \mid X = x', Z = z)}{\mathrm{P}(Y = y \mid X = x, Z = z)}$$

Make an Edit Proposal — Where to Edit?

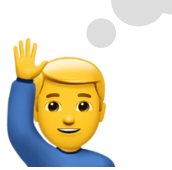
- **Conflict token detection**

Make an Edit Proposal — Where to Edit?

- **Conflict token detection**

$$P_{\text{cf}}(y_i^*) = \text{softmax}\left(\frac{P_{\text{LM}}(y_i^* \mid z, x, y_{<i}^*)}{P_{\text{LM}}(y_i^* \mid z, x', y_{<i}^*)}\right)$$

$$\text{CRR} = \frac{\text{P}(Y = y \mid X = x', Z = z)}{\text{P}(Y = y \mid X = x, Z = z)}$$



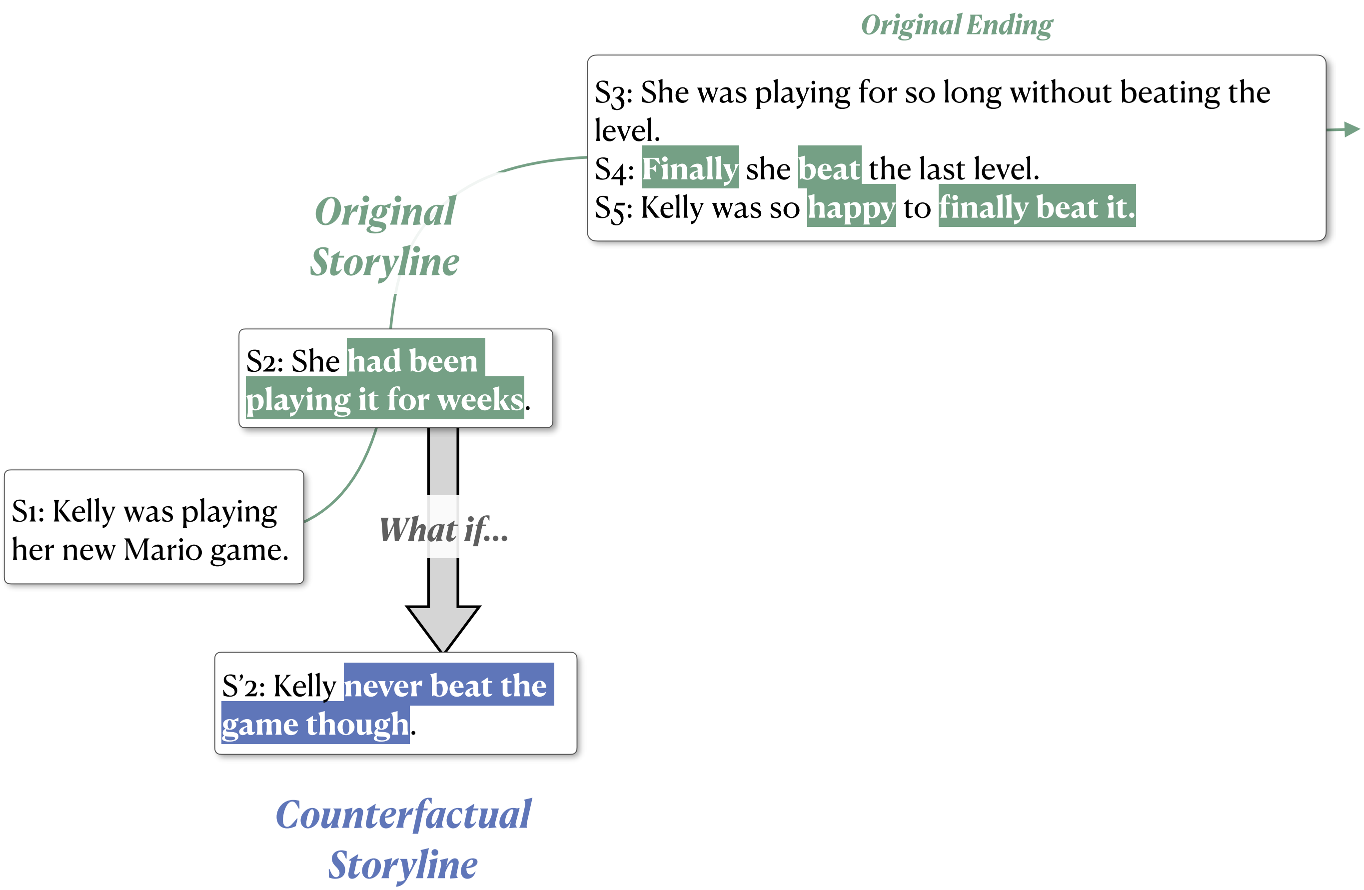
Make an Edit Proposal — Edit with What?

- **Modification actions**

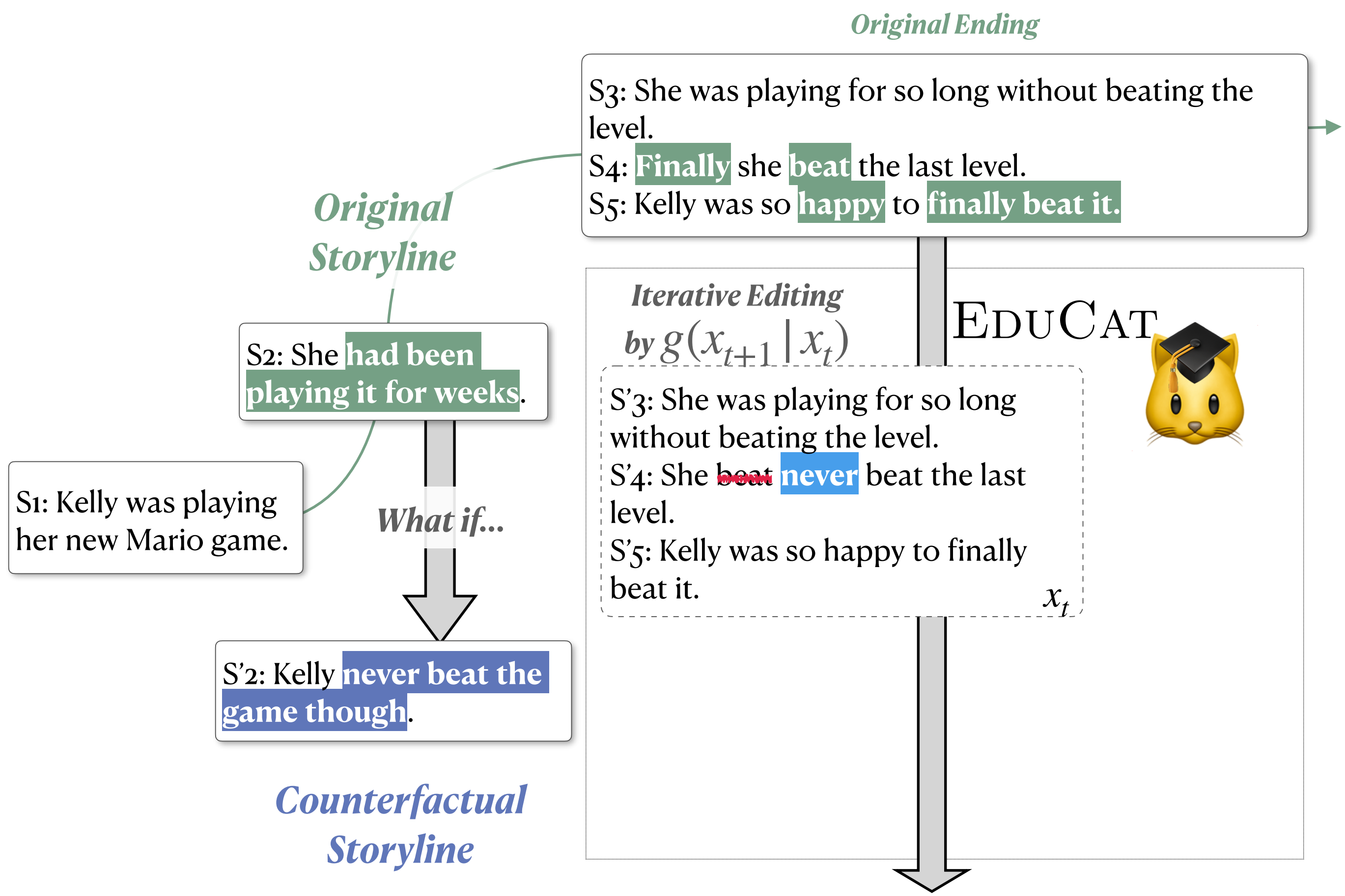
$$g(\mathrm{y}_{t+1} \mid \mathrm{y}_t) = \frac{1}{3} \sum_{\mathrm{op} \in \{\mathrm{r,d,i}\}} g_{\mathrm{op}}(\mathrm{y}_{t+1} \mid \mathrm{y}_t)$$

  - *Replace*: mask-predict with an MLM (e.g., BERT)
    - $g_r(\mathrm{y}_{t+1} \mid \mathrm{y}_t) = 1(w^c \in \mathcal{Q}) \cdot P_{\mathrm{MLM}}(w_m^* = w^c \mid \mathrm{x}_{-m})$
    - Sample from $P_{\mathrm{MLM}}(\,\cdot\,)$
  - *Insert*: insert a [MASK], then do *Replace*
  - *Delete*: reverse of *Insert*

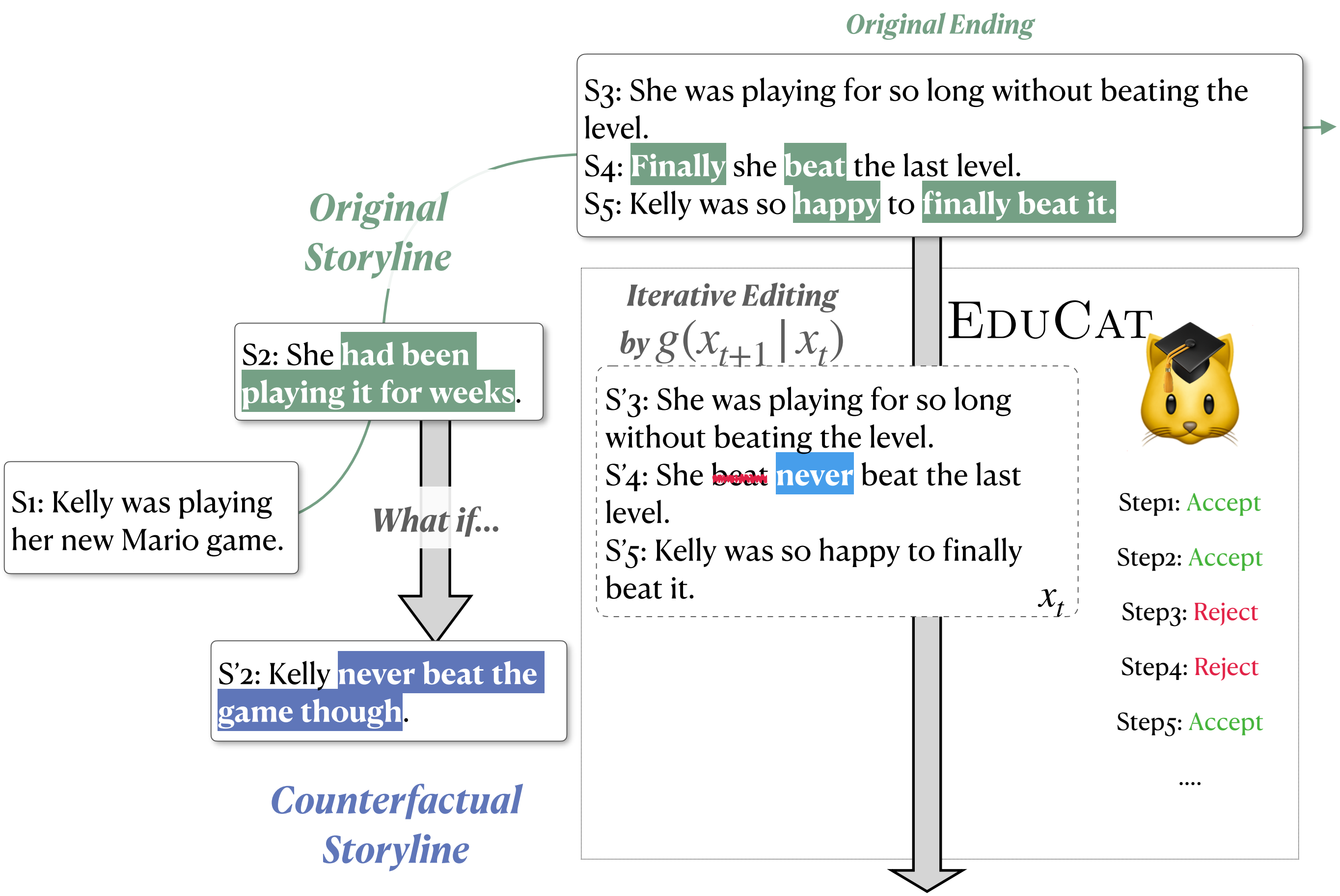
EDUCAT: Edit a Story Ending



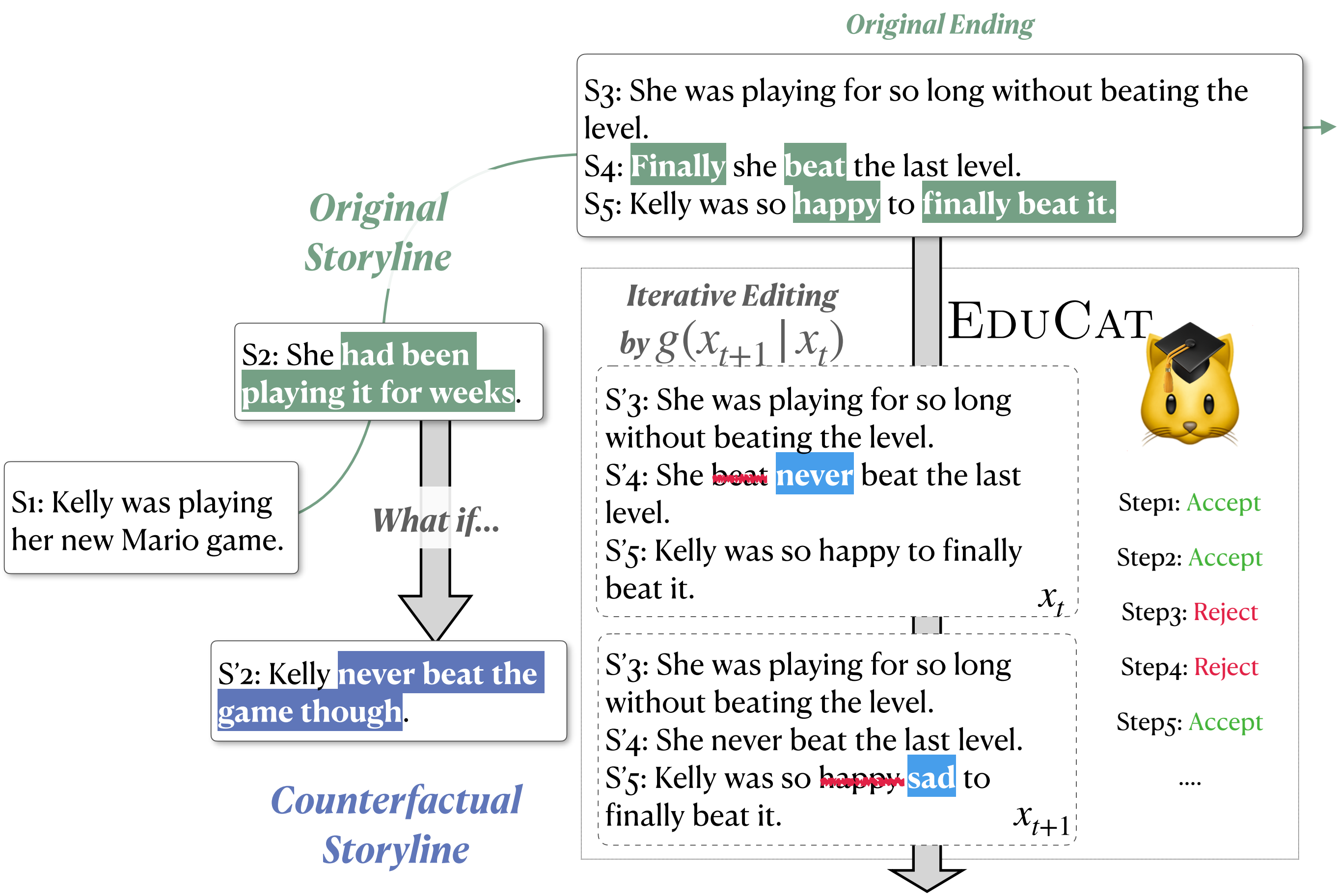
EDUCAT: Edit a Story Ending



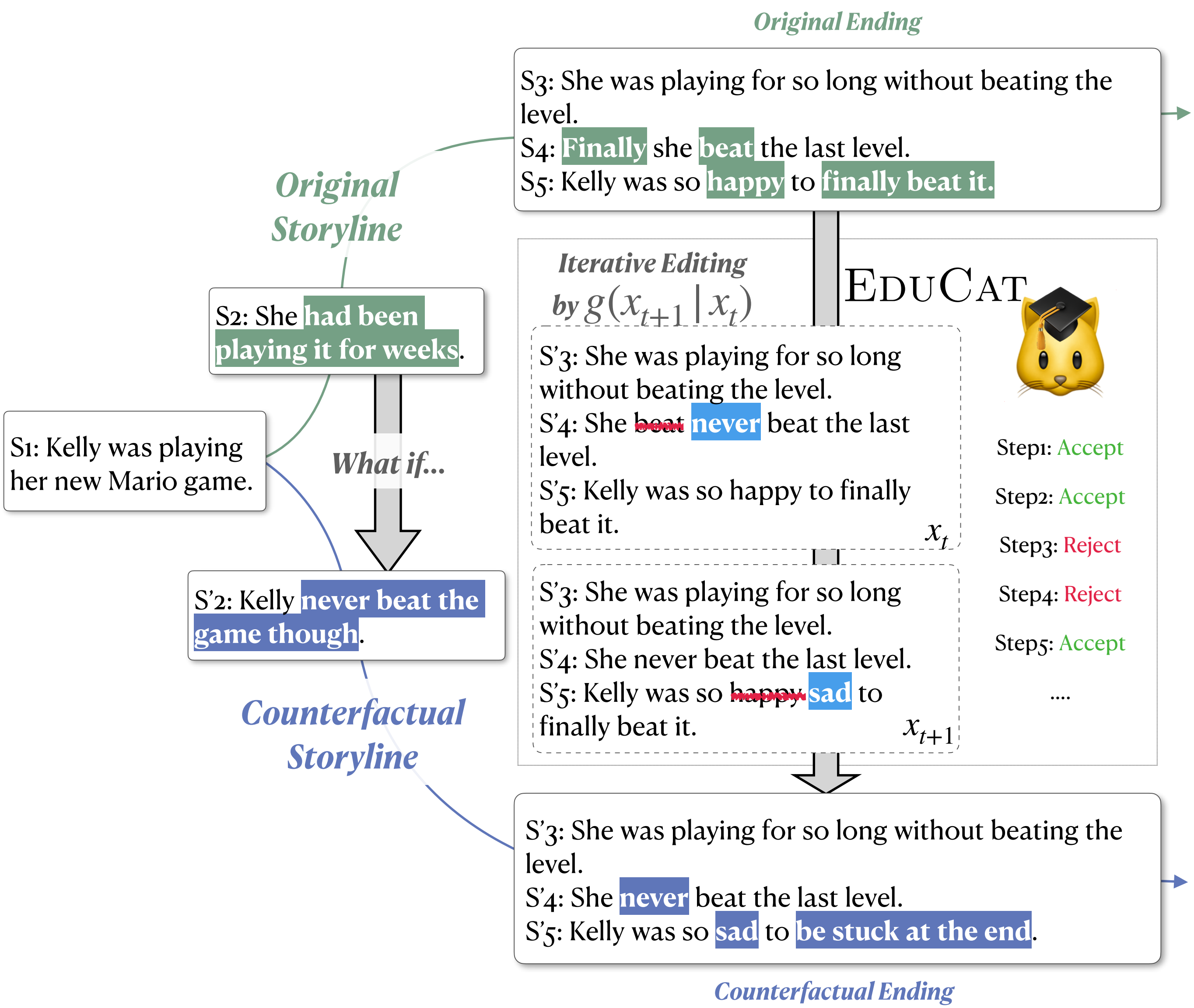
EDUCAT: Edit a Story Ending



EDUCAT: Edit a Story Ending



EDUCAT: Edit a Story Ending



Experiments: Dataset and Metrics

- **Dataset**
  - TimeTravel

	Train	Dev	Test
# counterfactual context (x')	96,867	1,871	1,871
# edited endings (y')	16,752	5,613	7,484

Table 1: Statistics of TIMETRAVEL dataset.

- **Metrics**
  - BLEU
  - BERTScore

Qin, Lianhui, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. **Counterfactual story reasoning and generation**. EMNLP 2019

Experiments: Dataset and Metrics

- **Dataset**
  - TimeTravel

	Train	Dev	Test
# counterfactual context (x')	96,867	1,871	1,871
# edited endings (y')	16,752	5,613	7,484

Table 1: Statistics of TIMETRAVEL dataset.

- **Metrics**
  - BLEU
  - BERTScore
  - **EntScore: a model-based discriminative metric**
    - Initial or counterfactual? Binary classification with RoBERTa
    - For coherence

Experiments: Dataset and Metrics

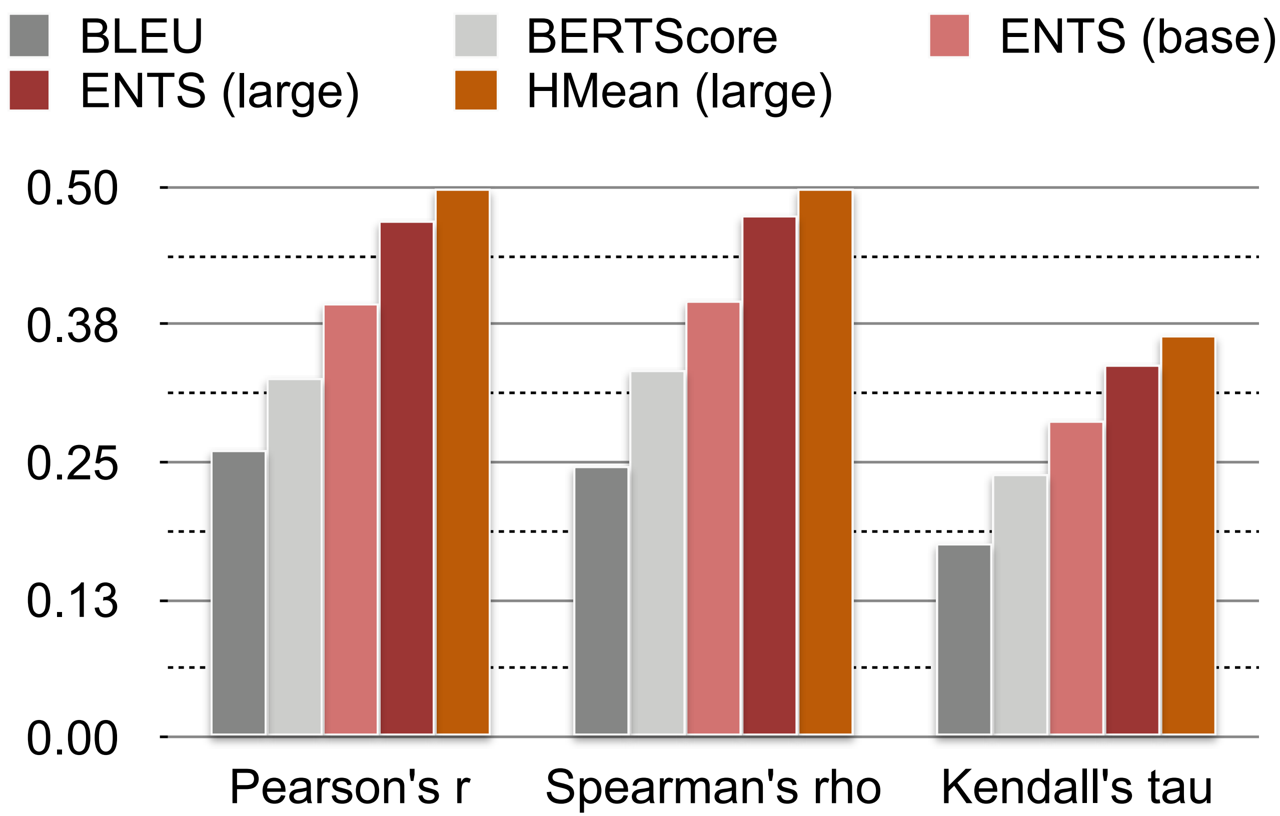
- **Dataset**
  - TimeTravel

	Train	Dev	Test
# counterfactual context (x')	96,867	1,871	1,871
# edited endings (y')	16,752	5,613	7,484

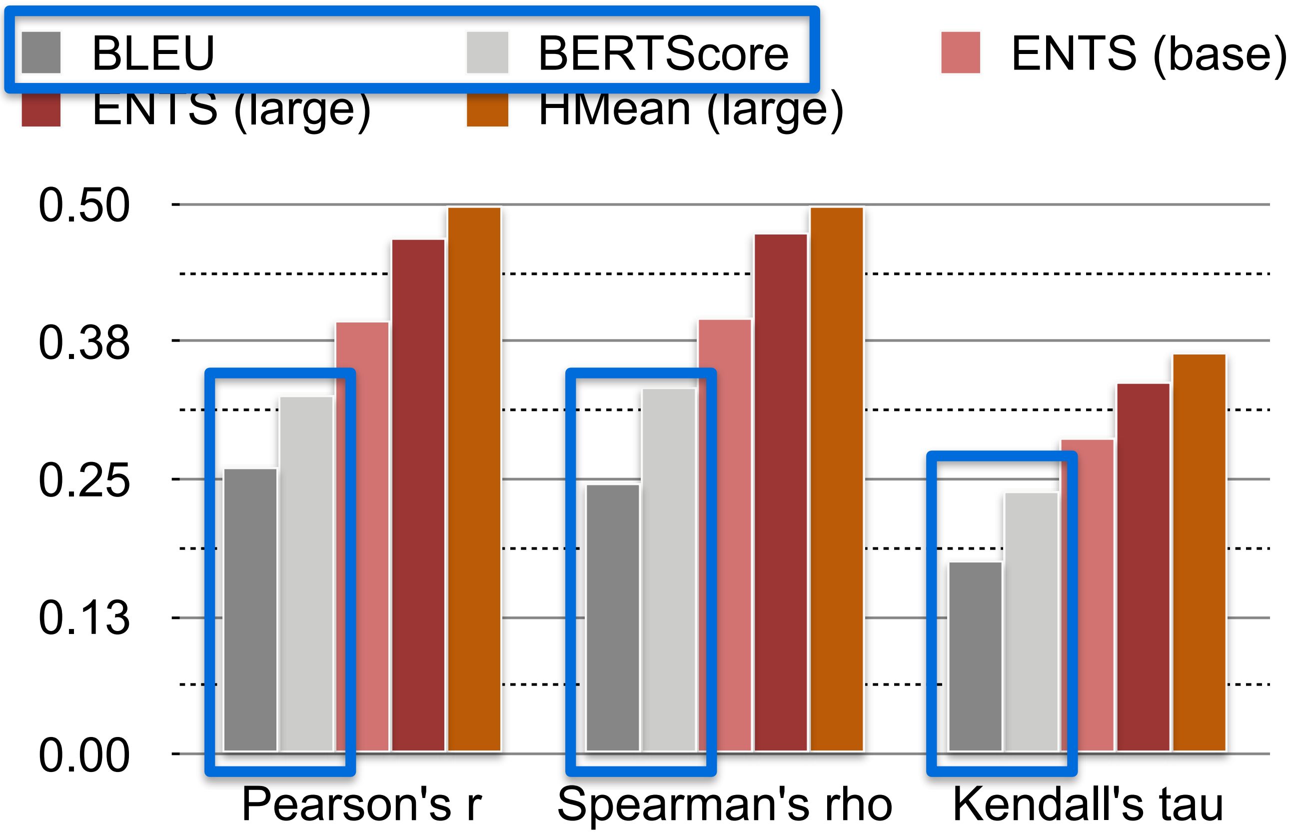
Table 1: Statistics of TIMETRAVEL dataset.

- **Metrics**
  - BLEU
  - BERTScore
  - **EntScore: a model-based discriminative metric**
    - Initial or counterfactual? Binary classification with RoBERTa
    - For coherence
  - **HMean: Harmonic Mean of EntScore and BLEU**
    - For the trade-off

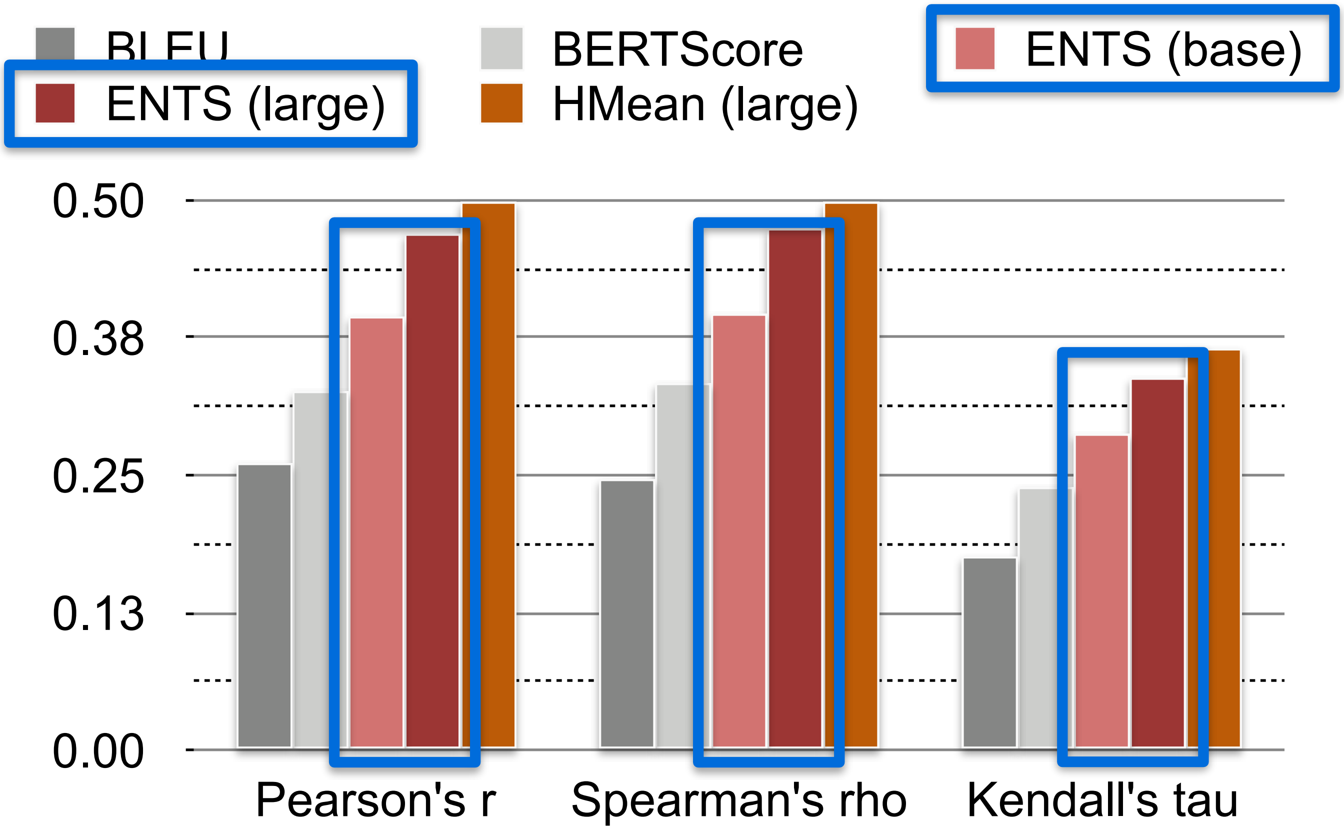
Quality of Metrics: Correlation with Humans



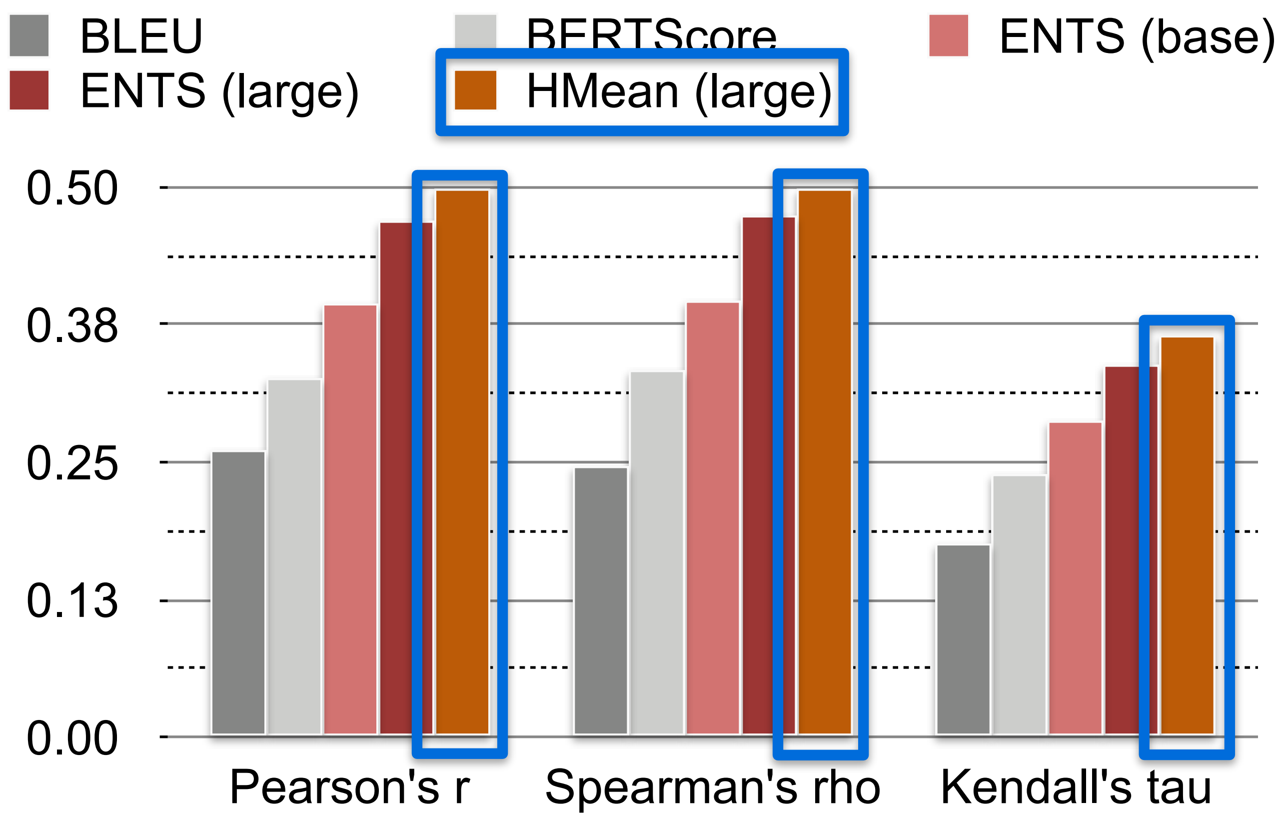
Quality of Metrics: Correlation with Humans



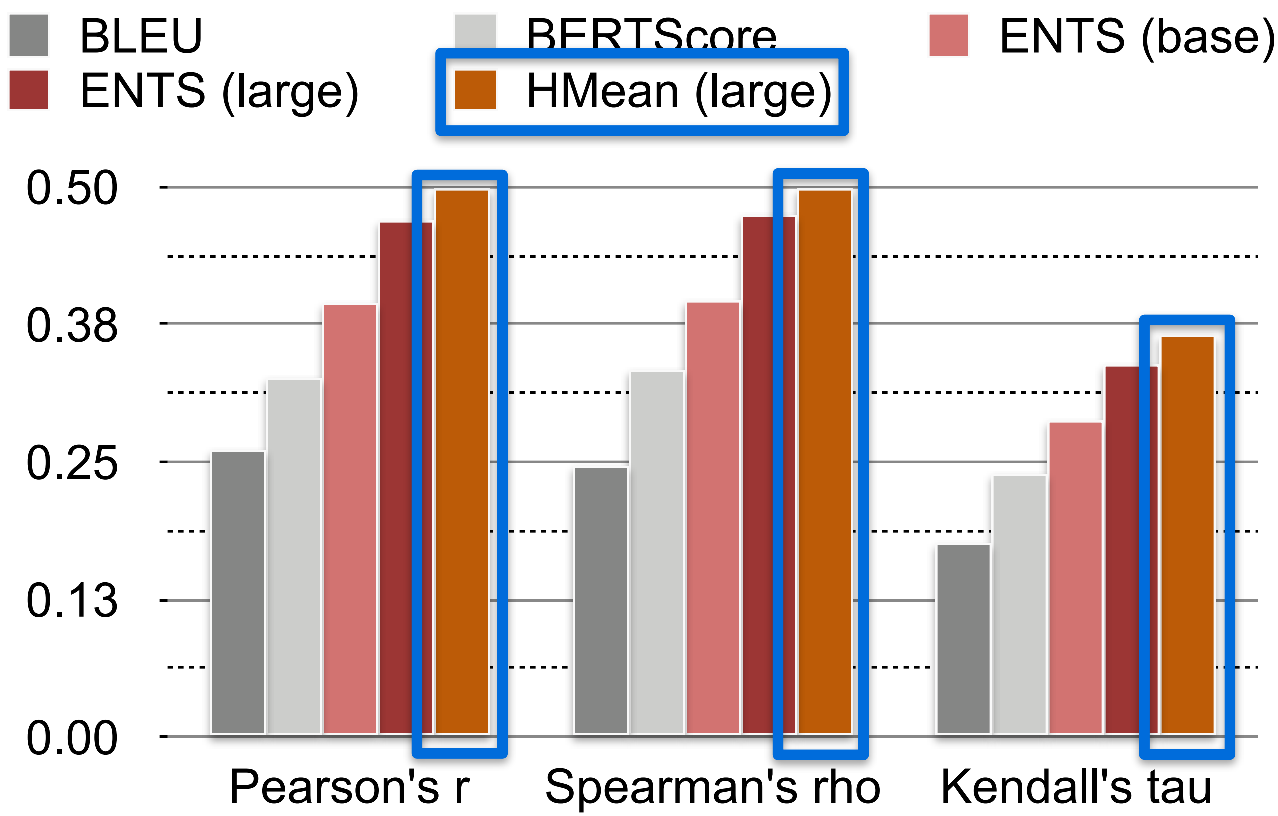
Quality of Metrics: Correlation with Humans



Quality of Metrics: Correlation with Humans



Quality of Metrics: Correlation with Humans



Better trade-off with HMean of ENTS and BLEU!

Automatic and Human Evaluation

Method	BLEU	BERT	ENTS$_l$	HMEAN
Supervised Training				
GPT-2$_M$ + SUP	76.35	81.72	35.06	48.05
Unsupervised Training				
GPT-2$_M$ + FT	3.90	53.00	52.77	7.26
Recon+CF	76.37	80.20	18.00	29.13
Off-the-shelf Pre-trained Models				
GPT-2$_M$	1.39	47.13	**54.21**	2.71
DELOREAN	23.89	59.88	51.40	32.62
CGMH	41.34	73.82	29.80	34.63
EDUCAT	**44.05**	**74.06**	32.28	**37.26**
Human	64.76	78.82	80.56	71.80

Table 3: Automatic evaluation results in the test set of TIME-TRAVEL. These methods use GPT-2$_M$ by default. ENTS$_l$ is short for ENTSCORE (large).

Automatic and Human Evaluation

Method	BLEU	BERT	ENTS$_l$	HMEAN
Supervised Training				
GPT-2$_M$ + SUP	76.35	81.72	35.06	48.05
Unsupervised Training				
GPT-2$_M$ + FT	3.90	53.00	52.77	7.26
Recon+CF	76.37	80.20	18.00	29.13
Off-the-shelf Pre-trained Models				
GPT-2$_M$	1.39	47.13	**54.21**	2.71
DELOREAN	23.89	59.88	51.40	32.62
CGMH	41.34	73.82	29.80	34.63
EDUCAT	**44.05**	**74.06**	32.28	**37.26**
Human	64.76	78.82	80.56	71.80

Table 3: Automatic evaluation results in the test set of TIME-TRAVEL. These methods use GPT-2$_M$ by default. ENTS$_l$ is short for ENTSCORE (large).

- EDUCAT is competitive against baselines but falls far behind humans.

Automatic and Human Evaluation

Method	BLEU	BERT	ENTS$_l$	HMEAN
		Supervised Training		
GPT-2$_M$ + SUP	76.35	81.72	35.06	48.05
		Unsupervised Training		
GPT-2$_M$ + FT	3.90	53.00	52.77	7.26
Recon+CF	76.37	80.20	18.00	29.13
		Off-the-shelf Pre-trained Models		
GPT-2$_M$	1.39	47.13	**54.21**	2.71
DELOREAN	23.89	59.88	51.40	32.62
CGMH	41.34	73.82	29.80	34.63
EDUCAT	**44.05**	**74.06**	32.28	**37.26**
Human	64.76	78.82	80.56	71.80

Table 3: Automatic evaluation results in the test set of TIME-TRAVEL. These methods use GPT-2$_M$ by default. ENTS$_l$ is short for ENTSCORE (large).

- EDUCAT is competitive against baselines but falls far behind humans.
- With massive edits, even a pre-trained GPT-2 can write coherent endings.

(Please check the paper for details.)

Automatic and Human Evaluation

Method	BLEU	BERT	ENTS$_l$	HMEAN
Supervised Training				
GPT-2$_M$ + SUP	76.35	81.72	35.06	48.05
Unsupervised Training				
GPT-2$_M$ + FT	3.90	53.00	52.77	7.26
Recon+CF	76.37	80.20	18.00	29.13
Off-the-shelf Pre-trained Models				
GPT-2$_M$	1.39	47.13	**54.21**	2.71
DELOREAN	23.89	59.88	51.40	32.62
CGMH	41.34	73.82	29.80	34.63
EDUCAT	**44.05**	**74.06**	32.28	**37.26**
Human	64.76	78.82	80.56	71.80

Table 3: Automatic evaluation results in the test set of TIME-TRAVEL. These methods use GPT-2$_M$ by default. ENTS$_l$ is short for ENTSCORE (large).

Methods	Coherence		
	Win	Tie	Lose
EDUCAT vs. DELOREAN	45%	32%	23%
EDUCAT vs. CGMH	32%	51%	17%
EDUCAT vs. Human	12%	24%	64%
	Min-edits		
EDUCAT vs. DELOREAN	64%	27%	9%
EDUCAT vs. CGMH	26%	49%	25%
EDUCAT vs. Human	16%	40%	44%

Table 4: Manual evaluation results, with scores denoting the percentage of *Win*, *Lose* or *Tie* when comparing EDUCAT with baselines.

- EDUCAT is competitive against baselines but falls far behind humans.
- With massive edits, even a pre-trained GPT-2 can write coherent endings.
- EDUCAT is competitive in coherence and minimal-edits under human evaluation.

(Please check the paper for details.)

Ablation Study

Ablation	BLEU	BERT	ENTS$_l$	HMEAN
EDUCAT (GPT-2$_S$)	39.82	72.35	31.72	35.31
EDUCAT (GPT-2$_M$)	44.05	74.06	32.28	37.26
– $\mathcal{X}_{\mathrm{Coh}}$	**44.20**	**74.27**	31.44	36.74
– *conflict detection*	40.96	73.61	30.79	35.16
– *both*	41.34	73.82	29.80	34.63
+ $\mathcal{X}_{\mathrm{Coh}}$ w/ ENTS$_b$	43.65	74.09	**42.03**	**42.83**

Table 5: Ablation study of EDUCAT in terms of conflict detection module and coherence score $\mathcal{X}_{\mathrm{Coh}}$. We also change the P_{Coh} in $\mathcal{X}_{\mathrm{Coh}}$ to the trained discriminative metric ENTSCORE.

- Both conflict detection and coherence objective work for the task.

(Please check the paper for details.)

Ablation Study

Ablation	BLEU	BERT	ENTS$_l$	HMEAN
EDUCAT (GPT-2$_S$)	39.82	72.35	31.72	35.31
EDUCAT (GPT-2$_M$)	44.05	74.06	32.28	37.26
– $\mathcal{X}_{\text{Coh}}$	**44.20**	**74.27**	31.44	36.74
– *conflict detection*	40.96	73.61	30.79	35.16
– *both*	41.34	73.82	29.80	34.63
+ $\mathcal{X}_{\text{Coh}}$ w/ ENTS$_b$	43.65	74.09	**42.03**	**42.83**

Table 5: Ablation study of EDUCAT in terms of conflict detection module and coherence score $\mathcal{X}_{\text{Coh}}$. We also change the P_{Coh} in $\mathcal{X}_{\text{Coh}}$ to the trained discriminative metric ENTSCORE.

- Both conflict detection and coherence objective work for the task.
- Can be further improved with a more sophisticated coherence checking model for $P_{\text{coh}}(\cdot)$ in $\mathcal{X}_{\text{coh}}(\cdot)$.

(Please check the paper for details.)

Case Study

S1: Gina had done everything she could think of to beat the heat.
S2: And it was only noon.
S3: The sun was still high in the sky.
S4: She decided she needed to go where there was air conditioning.
S5: She went inside a nearby cafe.

S’2: Luckily, it was evening and starting to cool off.

S’3: The sun had gotten lower in the sky.
S’4: She decided next time it was so hot she needed to go where there was air conditioning.
S’5: So she planned to go inside a nearby cafe.

HUMAN

Case Study

S1: Gina had done everything she could think of to beat the heat.
S2: And it was only noon.
S3: The sun was still high in the sky.
S4: She decided she needed to go where there was air conditioning.
S5: She went inside a nearby cafe.

S'2: Luckily, it was evening and starting to cool off.

S'3: The sun had gotten lower in the sky.
S'4: She decided next time it was so hot she needed to go where there was air conditioning.
S'5: So she planned to go inside a nearby cafe.

HUMAN

S'3: The sun was still high in the sky.
S'4: She decided she needed to go outside and get some fresh air.
S'5: She went inside and got some fresh air.

DELOREAN

Case Study

S1: Gina had done everything she could think of to beat the heat.
S2: And it was only noon.
S3: The sun was still high in the sky.
S4: She decided she needed to go where there was air conditioning.
S5: She went inside a nearby cafe.

S'2: Luckily, it was evening and starting to cool off.

S'3: The sun had gotten lower in the sky.
S'4: She decided next time it was so hot she needed to go where there was air conditioning.
S'5: So she planned to go inside a nearby cafe.

HUMAN

S'3: The sun was still high in the sky.
S'4: She decided she needed to go outside and get some fresh air.
S'5: She went inside and got some fresh air.

DELOREAN

S'3: The sun was high in the sky.
S'4: She decided she needed to go somewhere where there was air.
S'5: She went to the beach.

CGMH

Case Study

S1: Gina had done everything she could think of to beat the heat.
S2: And it was only noon.
S3: The sun was still high in the sky.
S4: She decided she needed to go where there was air conditioning.
S5: She went inside a nearby cafe.

S'2: Luckily, it was evening and starting to cool off.

S'3: The sun had gotten lower in the sky.
S'4: She decided next time it was so hot she needed to go where there was air conditioning.
S'5: So she planned to go inside a nearby cafe.

HUMAN

S'3: The sun was still high in the sky.
S'4: She decided she needed to go outside and get some fresh air.
S'5: She went inside and got some fresh air.

DELOREAN

S'3: The sun was high in the sky.
S'4: She decided she needed to go somewhere where there was air.
S'5: She went to the beach.

CGMH

S'3: The sun was low in the sky.
S'4: She decided that she needed to go somewhere where there was no air conditioning.
S'5: She headed to the park.

EDUCAT

Takeaways

- Editing-based methods are also well-suited for generative counterfactual reasoning.
- Better components for EDUCAT will benefit unsupervised story rewriting.
  - e.g., better coherence score, more desired properties, etc.
- Better metrics should be studied for evaluating this task!

Have Fun with EDUCAT!

Checkout our code at **GitHub**!

https://github.com/jiangjiechen/EDUCAT