Converge to the Truth: Factual Error Correction via Iterative Constrained Editing

Jiangjie Chen, Rui Xu, Wenxuan Zeng,
Changzhi Sun, Lei Li, Yanghua Xiao

FUDAN UNIVERSITY
University of Electronic Science and Technology of China
ByteDance
UC SANTA BARBARA

Grady Booch ✓
@Grady_Booch

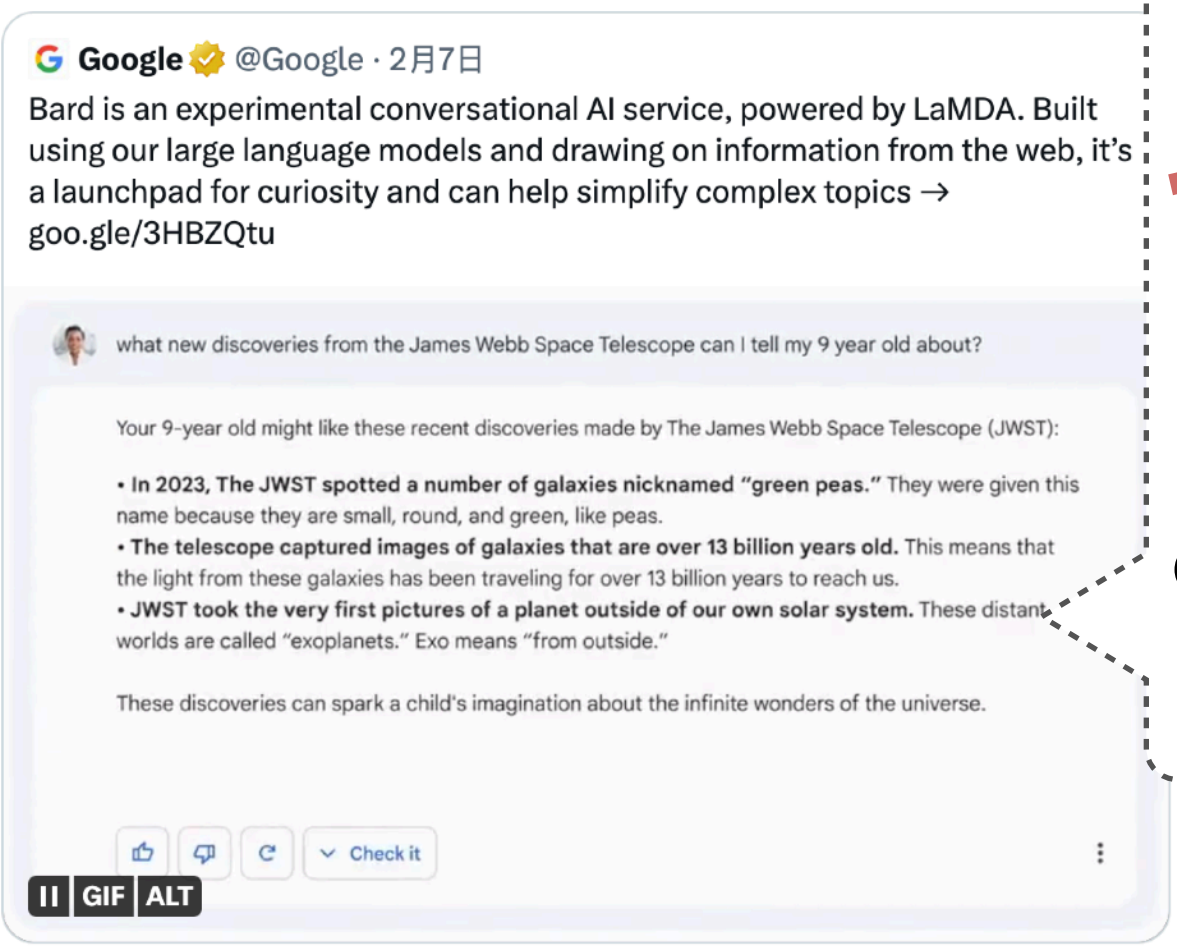
Misinformation at scale; bullshit as a service.

(The European Very Large Telescope- not the JWST - took the first optical photograph of an exoplanet in 2004.)

翻译推文

Google ✓ @Google · 2月7日

Bard is an experimental conversational AI service, powered by LaMDA. Built using our large language models and drawing on information from the web, it's a launchpad for curiosity and can help simplify complex topics → goo.gle/3HBZQtu



European Very Large Telescope

~~James Webb Space~~ Telescope took the very first pictures of a planet outside of our own solar system.

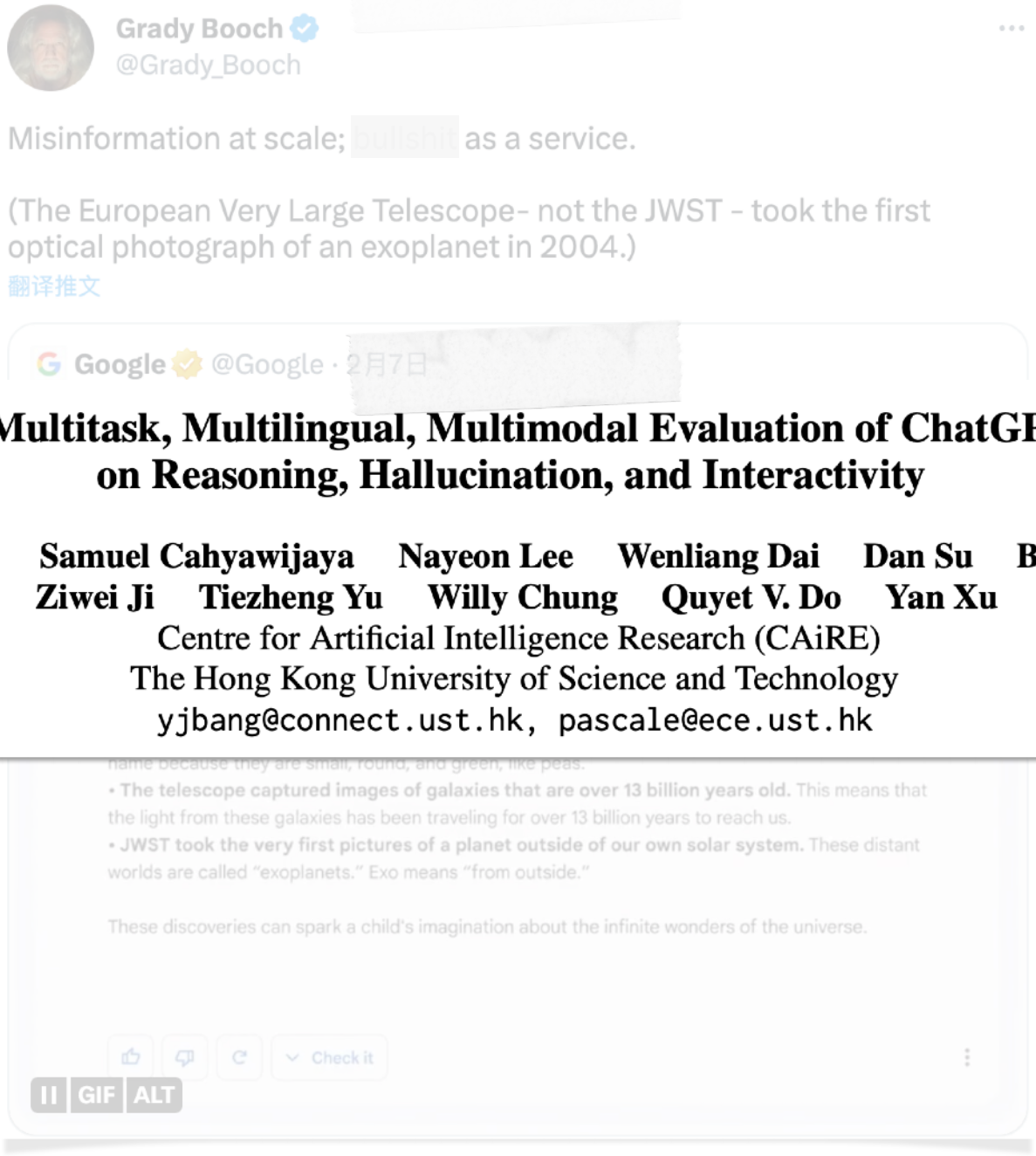
TIME

TECH • SOCIAL MEDIA

A Factual Error by Bard AI Chatbot Just Cost Google $100 Billion

February 9 2023

https://time.com/6254226/alphabet-google-bard-100-billion-ai-error/



A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity

Yejin Bang **Samuel Cahyawijaya** **Nayeon Lee** **Wenliang Dai** **Dan Su** **Bryan Wilie**
Holy Lovenia **Ziwei Ji** **Tiezheng Yu** **Willy Chung** **Quyet V. Do** **Yan Xu** **Pascale Fung**
Centre for Artificial Intelligence Research (CAiRE)
The Hong Kong University of Science and Technology
yjbang@connect.ust.hk, pascale@ece.ust.hk

Factual Error Correction

with minimal modification

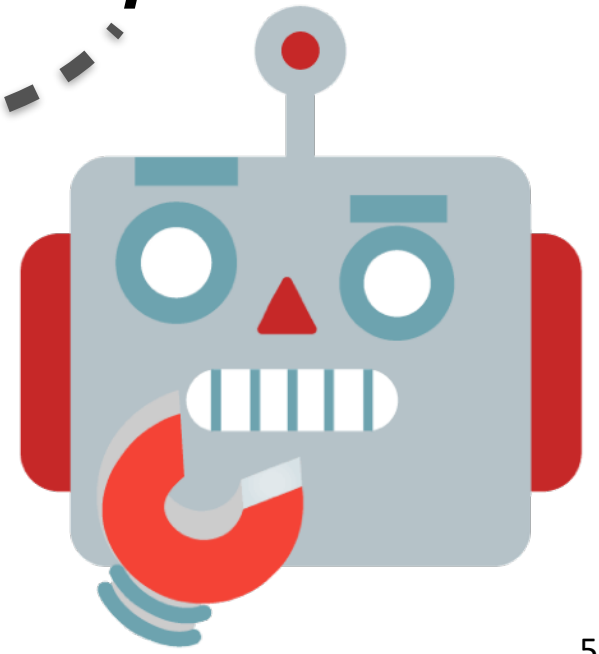
Claim

Will Smith starred in **The Truman Show** in **2006**.

Wikipedia

Retrieved Evidence

The Pursuit of Happyness is a **2006** American biographical drama film directed by Gabriele Muccino and starring **Will Smith** as Chris Gardner, a homeless …



Will Smith starred in **The Pursuit for Happyness** in **2006**.

Corrected Claim

How to Correct Claims?

💡**Previous**: one-pass mask-then-correct generation

Will Smith starred in The Truman Show in 2006. ||  *Retrieved Evidence*

Mask spans

Will Smith starred in [MASK] [MASK] [MASK] in 2006. || 

LM generation

Will Smith starred in **<u>The Pursuit for Happiness</u>** Show in 2006.

No supervised data?
using Fact verification model to rank!

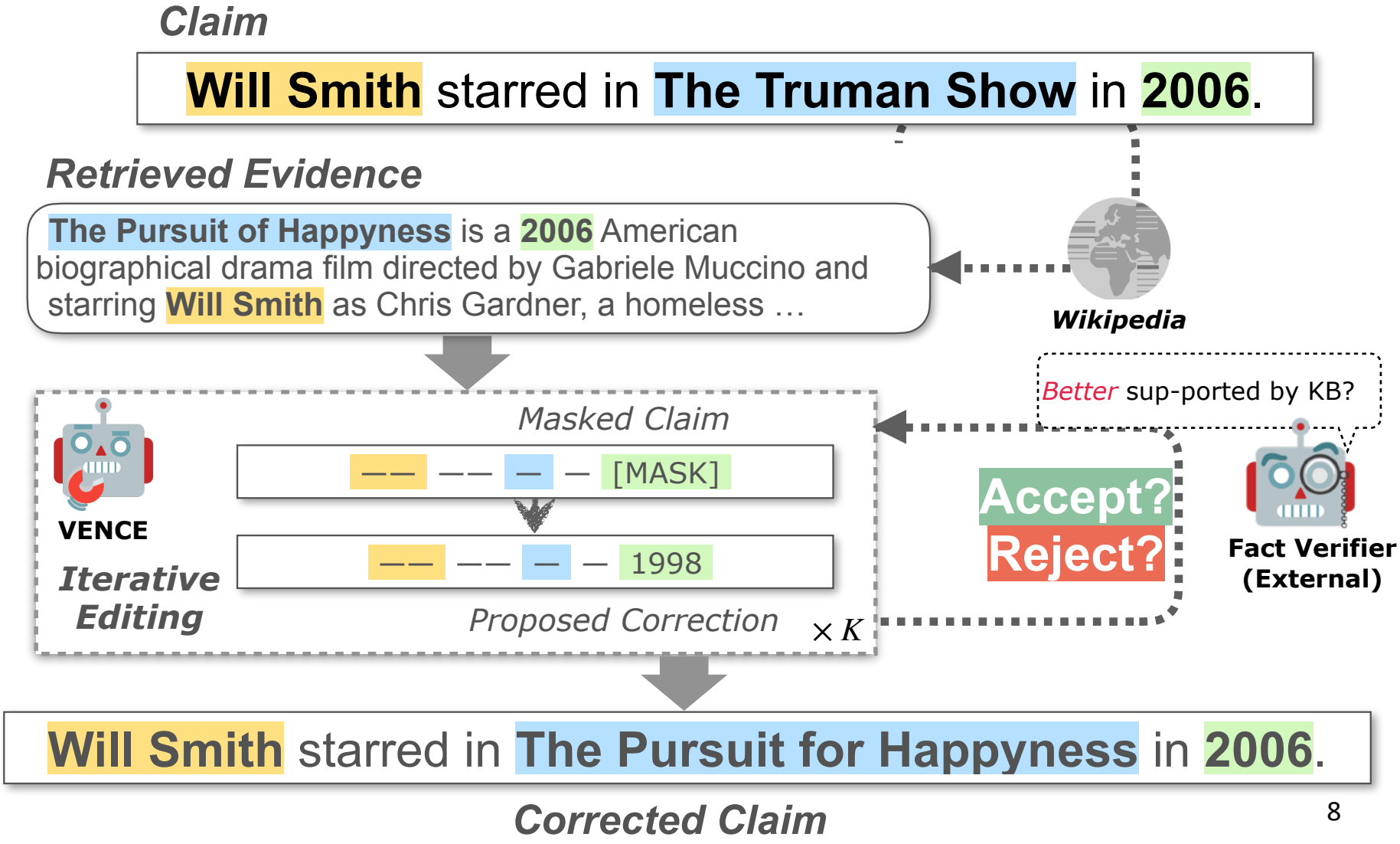
James Thorne, Andreas Vlachos, **Evidence-based Factual Error Correction.** ACL 2021.

Challenges

- **Semantic Drift of Intended Meaning**
  - "New York has 18 million population." => "New York is in the US."
  - ✳ **=> *enforcing minimal-edits***

- **Lack Annotated Dataset** for FEC
  - High-quality datasets are costly to build.
  - Most datasets are synthetic.
  - ✳ **=> *Self-supervised editing proposal***

VENCE: VErificatioN-guided Constrained Editing



Desired Properties

Sampling from Target distribution

$$\pi(x) = \frac{e^{-\mathscr{E}(x)}}{Z}$$

$$\mathscr{E}(x) = \mathscr{E}_{\mathrm{LM}}(x) + \mathscr{E}_{\mathrm{V}}(x) + \mathscr{E}_{\mathrm{H}}(x)$$

Fluency:
- Language Modeling

Truthfulness:
- Fact Verification

Num-edits:
- Hamming Distar

$$\mathscr{E}_{\mathrm{LM}}(x) = -\sum_i \log P_{\mathrm{MLM}}(w_i \,|\, x_{-i})$$

$$\mathscr{E}_{\mathrm{V}}(x) = -\log P_{\mathrm{V}}(\mathrm{Supported} \,|\, x, E)$$

$$\mathscr{E}_{\mathrm{H}}(x) = \mathrm{HammingDistance}(x, x^0)$$

VENCE

Input: x

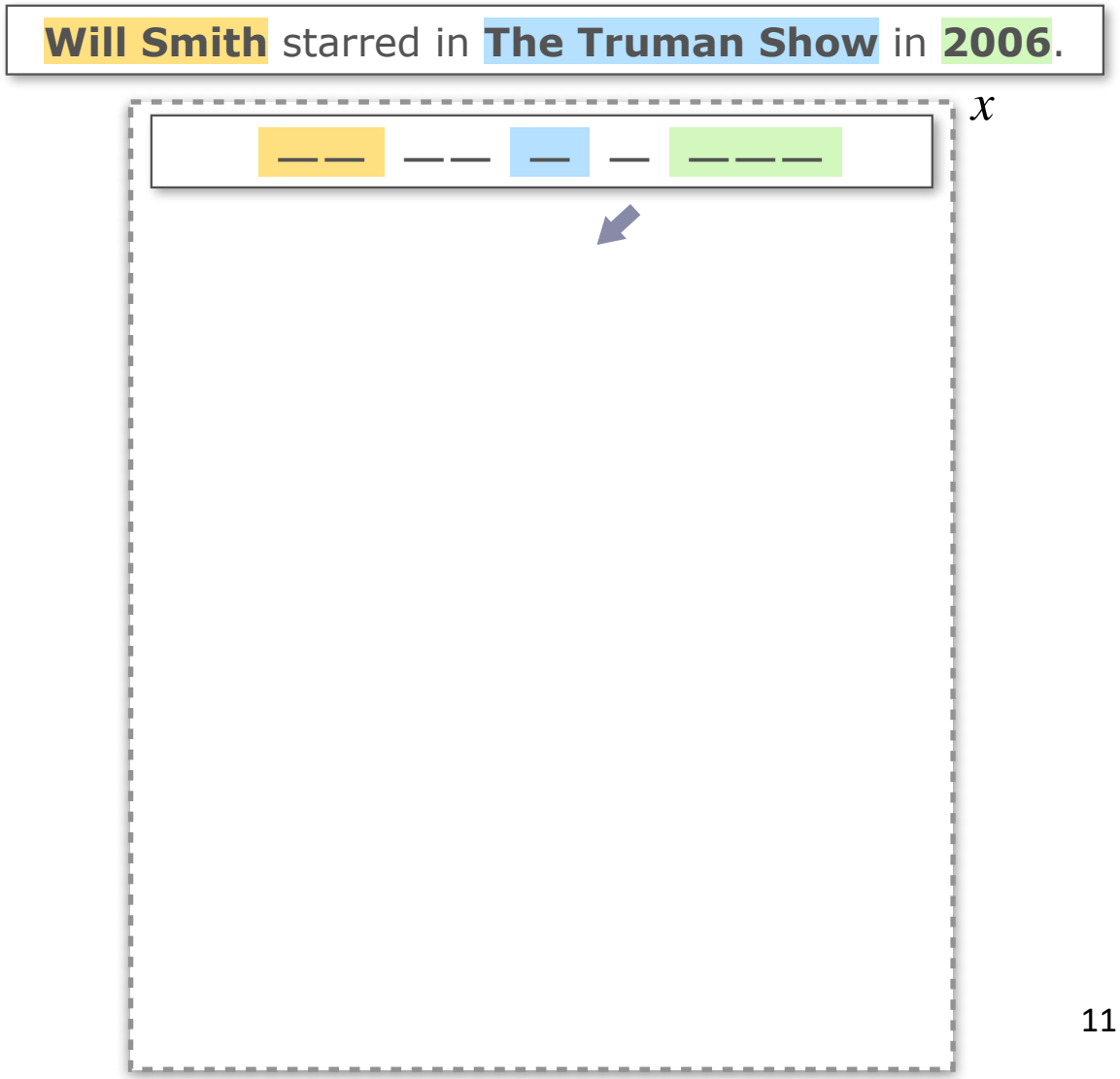
Will Smith starred in **The Truman Show** in **2006**.

VENCE

Input: x

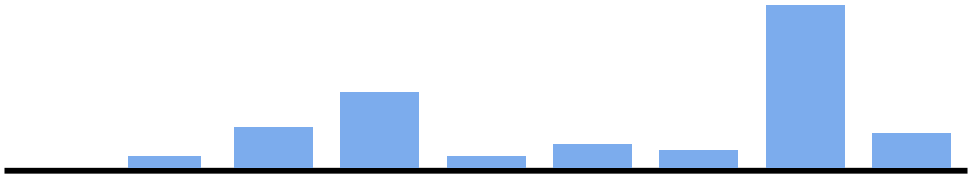
↓

Randomly select a position to edit



Editing Position Proposal: $P_1(m \mid x)$

- Sample a position based on Verifier (P_V)'s **normalized gradient distribution**.



Normalized gradient distribution over tokens

- **Multi-token Entity Masking**
  - ✓ Will Smith starred in [MASK] in 2006.
  - – ~~Will Smith starred in The [MASK] Show in 2006.~~

VENCE

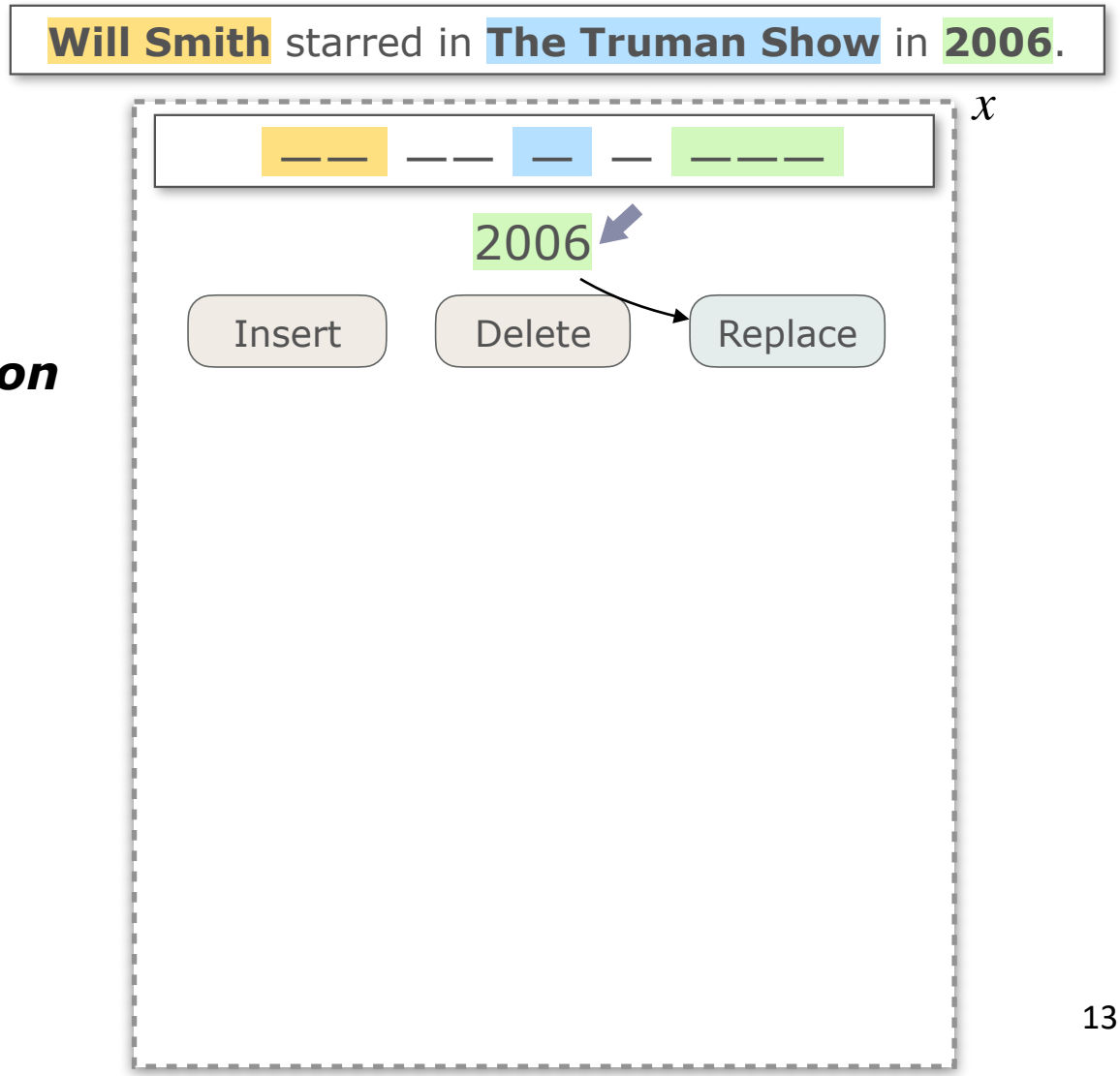
Input: x

↓

Randomly select a position to edit

↓

Choose an editing action



VENCE

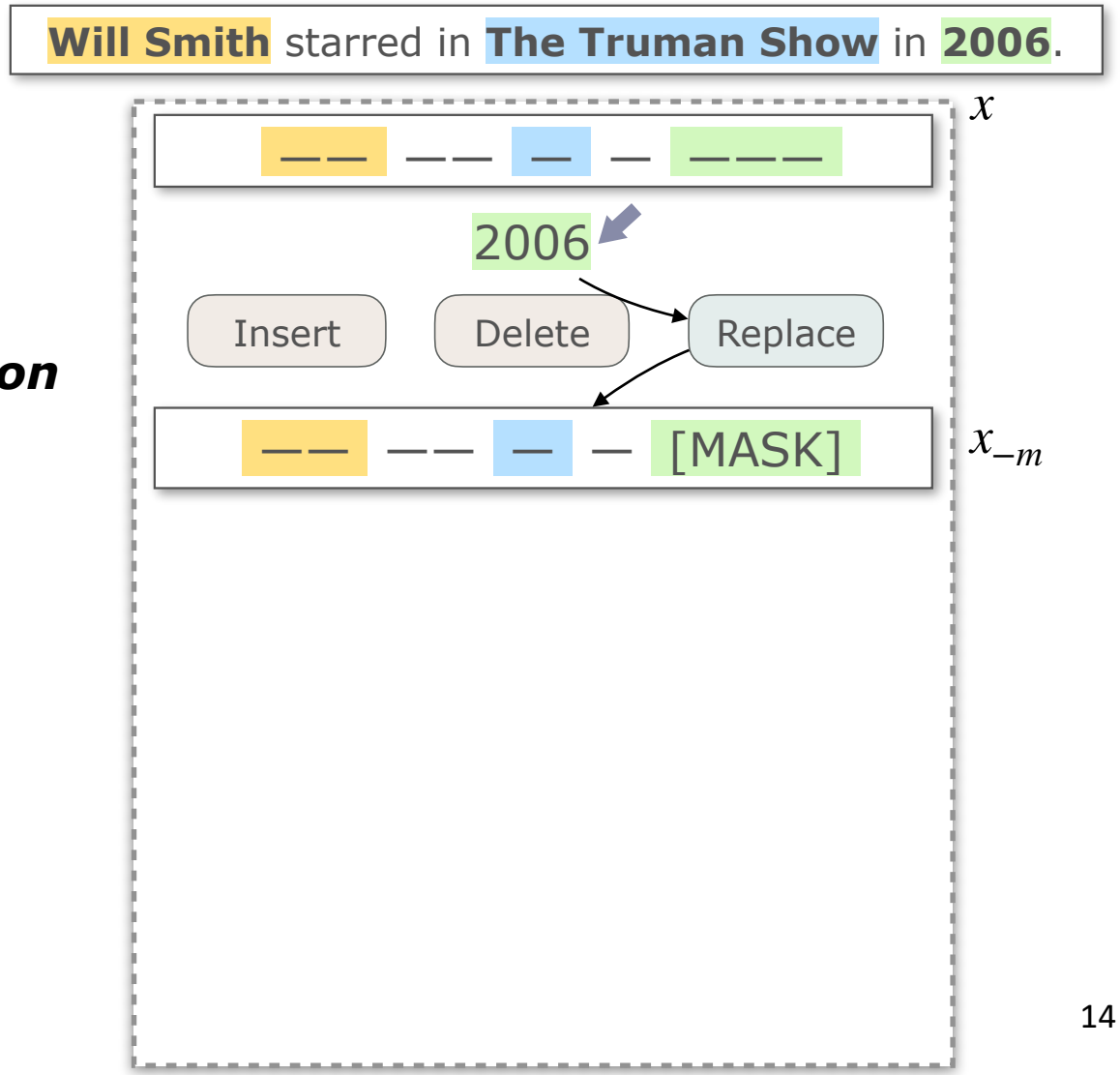
Input: x

↓

Sample a position to edit with a verifier

↓

Choose an editing action



VENCE

Input: x

↓

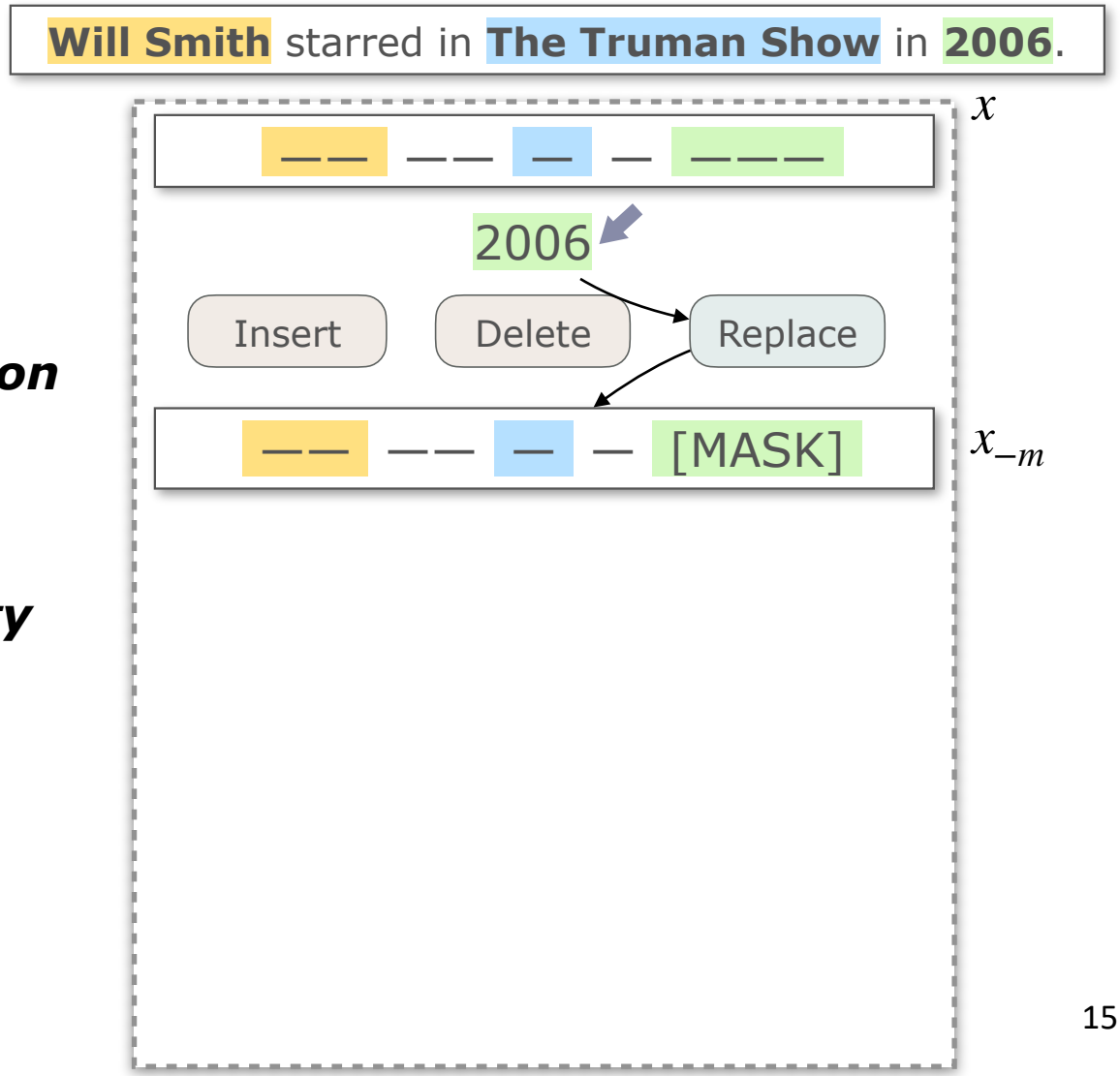
Sample a position to edit with a verifier

↓

Choose an editing action

↓

Sample a token/entity

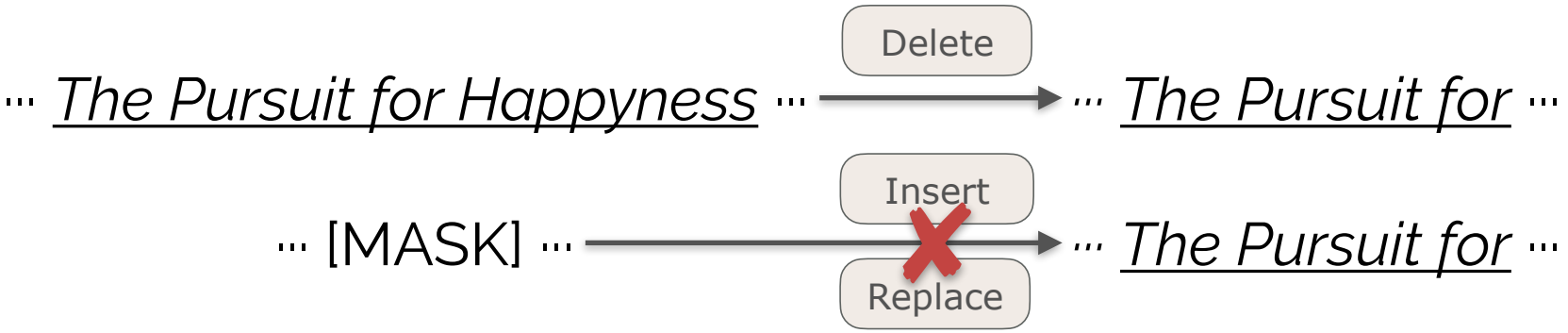


Challenges brought by Multi-token Entity Masking

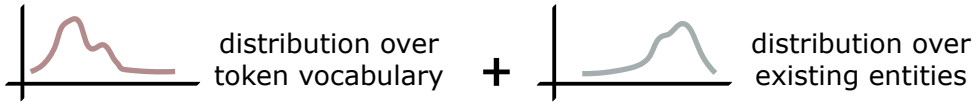
- The Markov chain must satisfy ***detailed balance condition*** for the sampling to converge on $\pi(x)$.

$$\pi(x')g(x \mid x') = \pi(x)g(x' \mid x)$$

- ***Price for editing entities***: predicting a list of tokens as entities given one mask is not reversible.



Solution

- **Separating** the sampling space into a token space and an entity space
  - 

- **Generative proposal** with a T5 model
- **Reversible**
  - A delete+insert action combination can communicate two spaces.
  - e.g., delete(entity) + insert(token)

Generative Proposal Model $P_3(x' \mid x_{-m}, a)$: Token vs. Entity

- **Replacement**

$$P_3(x' \mid x_{-m}, \texttt{rep}) = \begin{cases} P_3^{\text{ent}}(x' \mid x_{-m}, \texttt{rep}), & x_m = \texttt{ent} \\ P_3^{\text{tok}}(x' \mid x_{-m}, \texttt{rep}), & x_m = \texttt{tok} \end{cases}$$

- **Insertion**

$$P_3(x' \mid x_{-m}, \texttt{ins}) = \alpha P_3^{\text{tok}}(x' \mid x_{-m}, \texttt{ins}) + (1-\alpha) P_3^{\text{ent}}(x' \mid x_{-m}, \texttt{ins})$$

- **Deletion** (reverse of insertion)

$$P_3(x' \mid x_{-m}, \texttt{del}) = 1$$

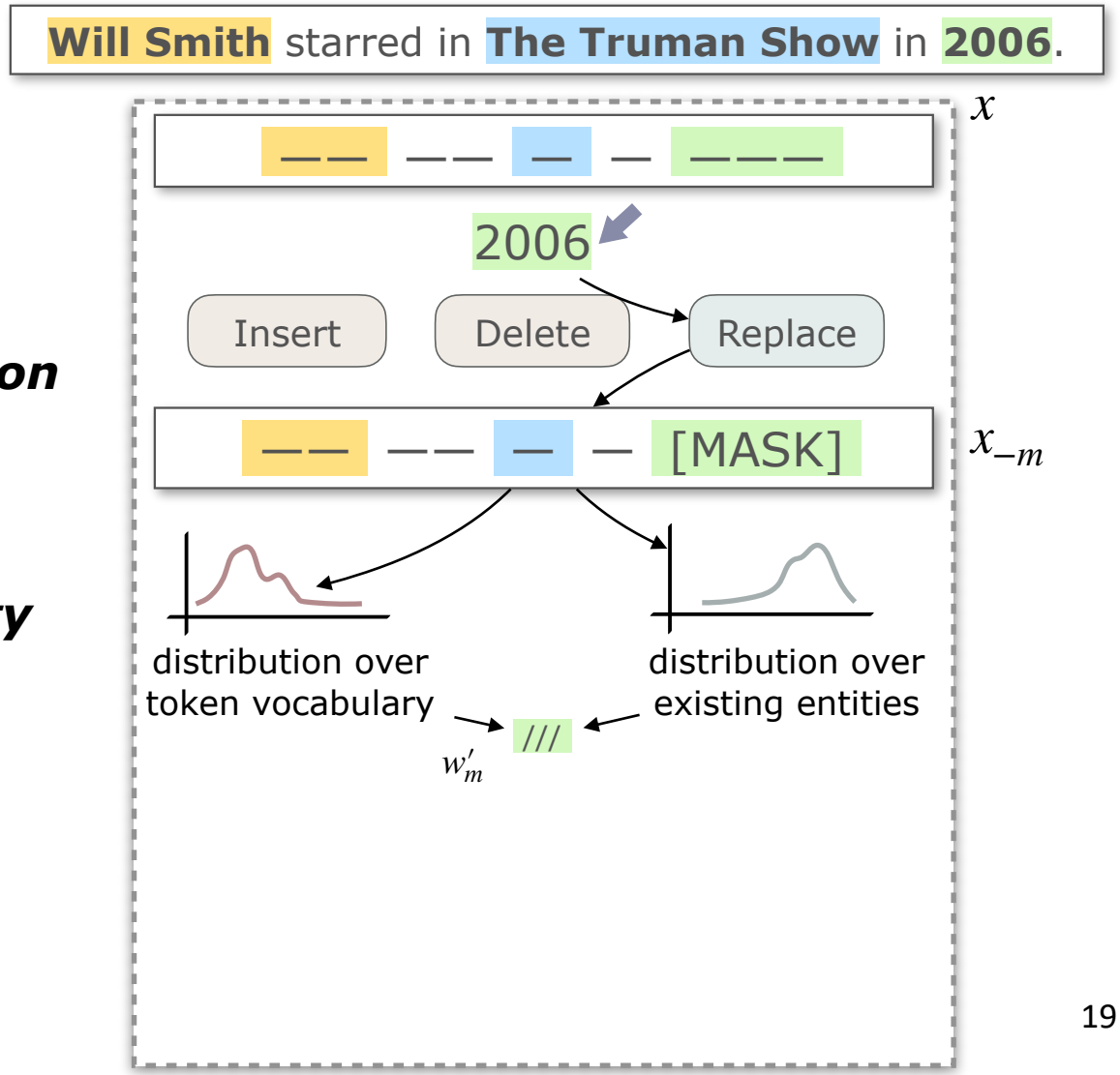
VENCE

Input: x

Sample a position to edit with a verifier

Choose an editing action

Sample a token/entity



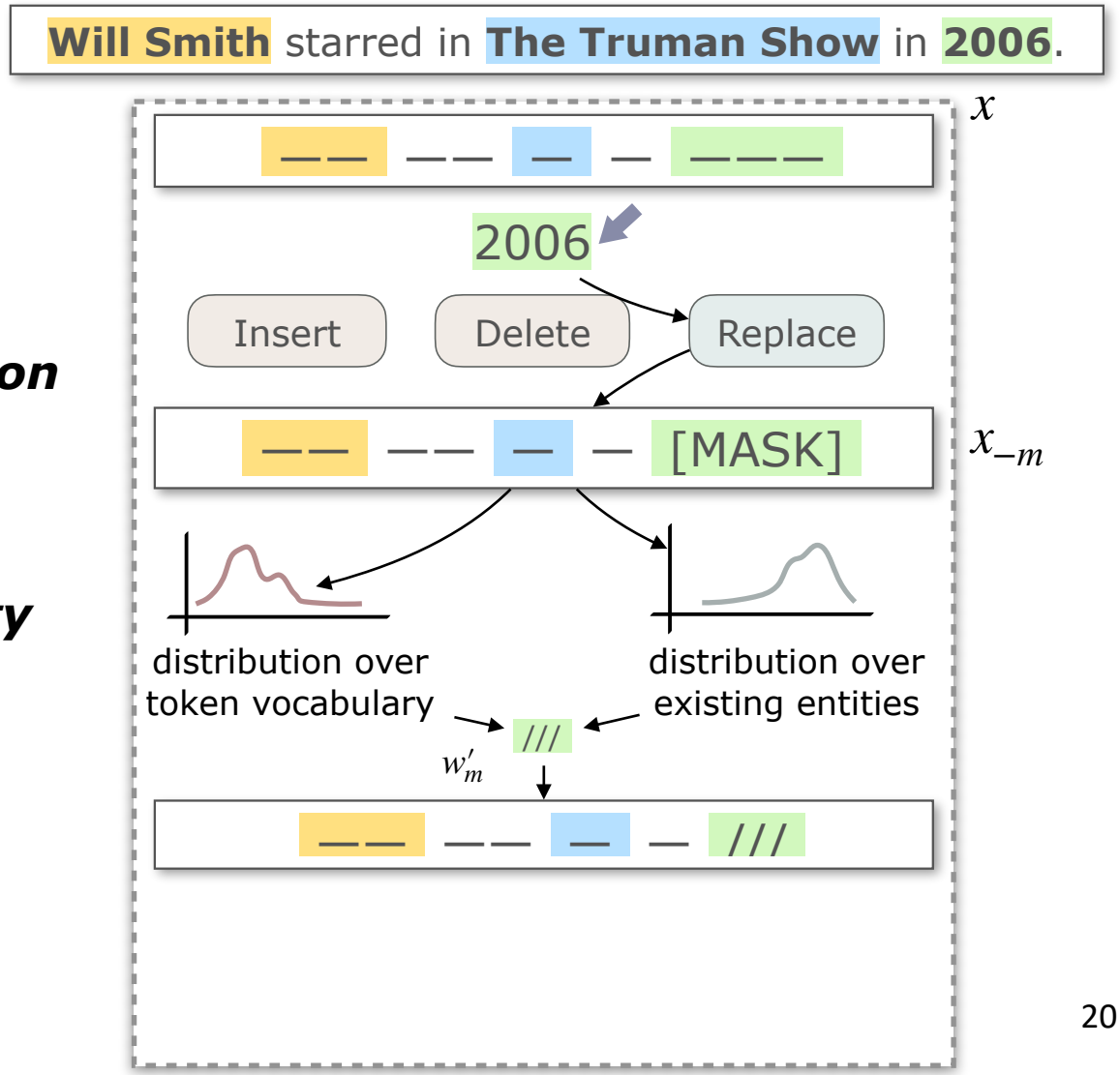
VENCE

Input: x

Sample a position to edit with a verifier

Choose an editing action

Sample a token/entity



VENCE

Input: x

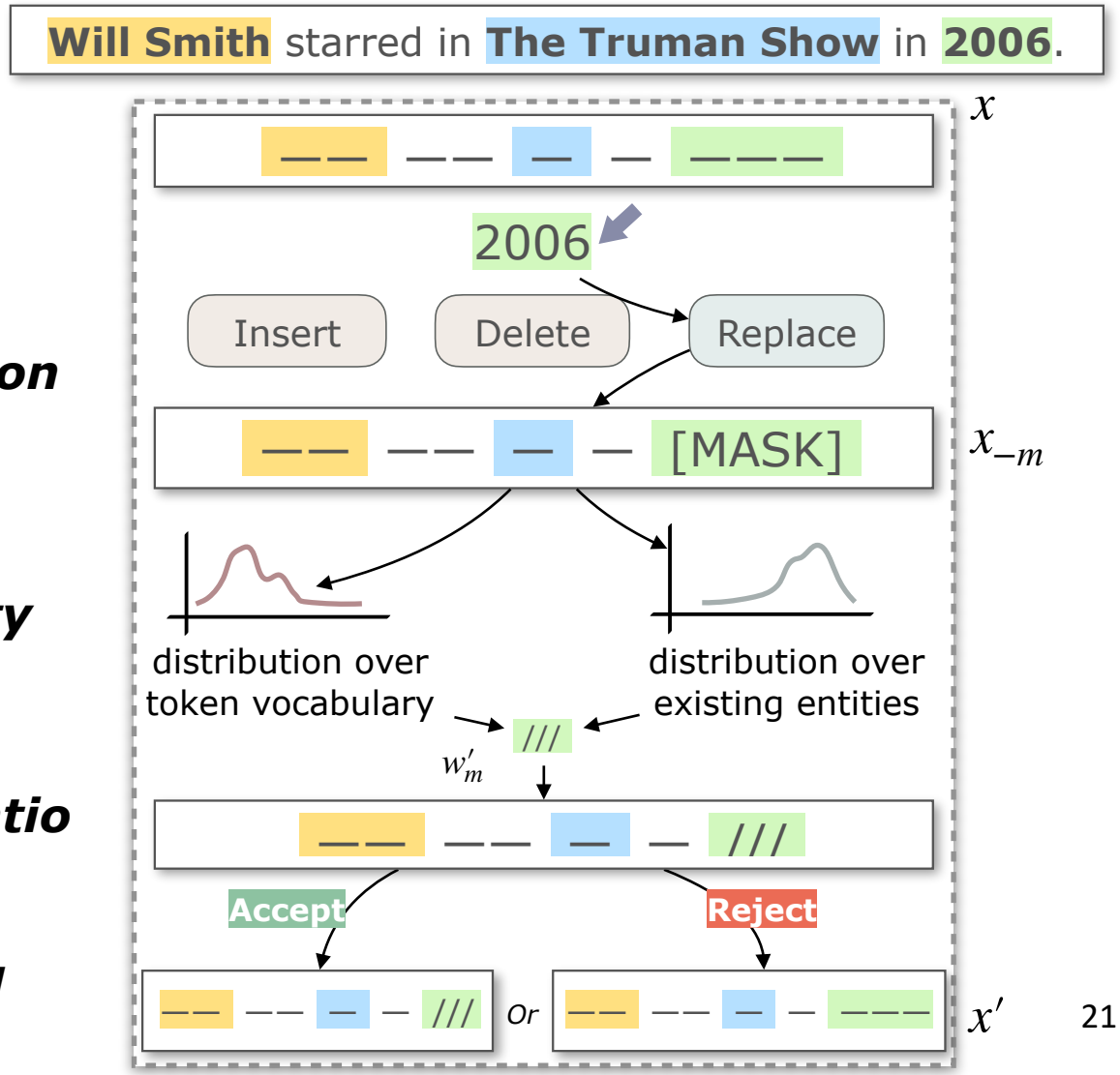
Sample a position to edit with a verifier

Choose an editing action

Sample a token/entity

Compute acceptance ratio with Energy Functions

Update x' ***if accepted***



VENCE

Input: x

↓

Sample a position to edit with a verifier

↓

Choose an editing action

↓

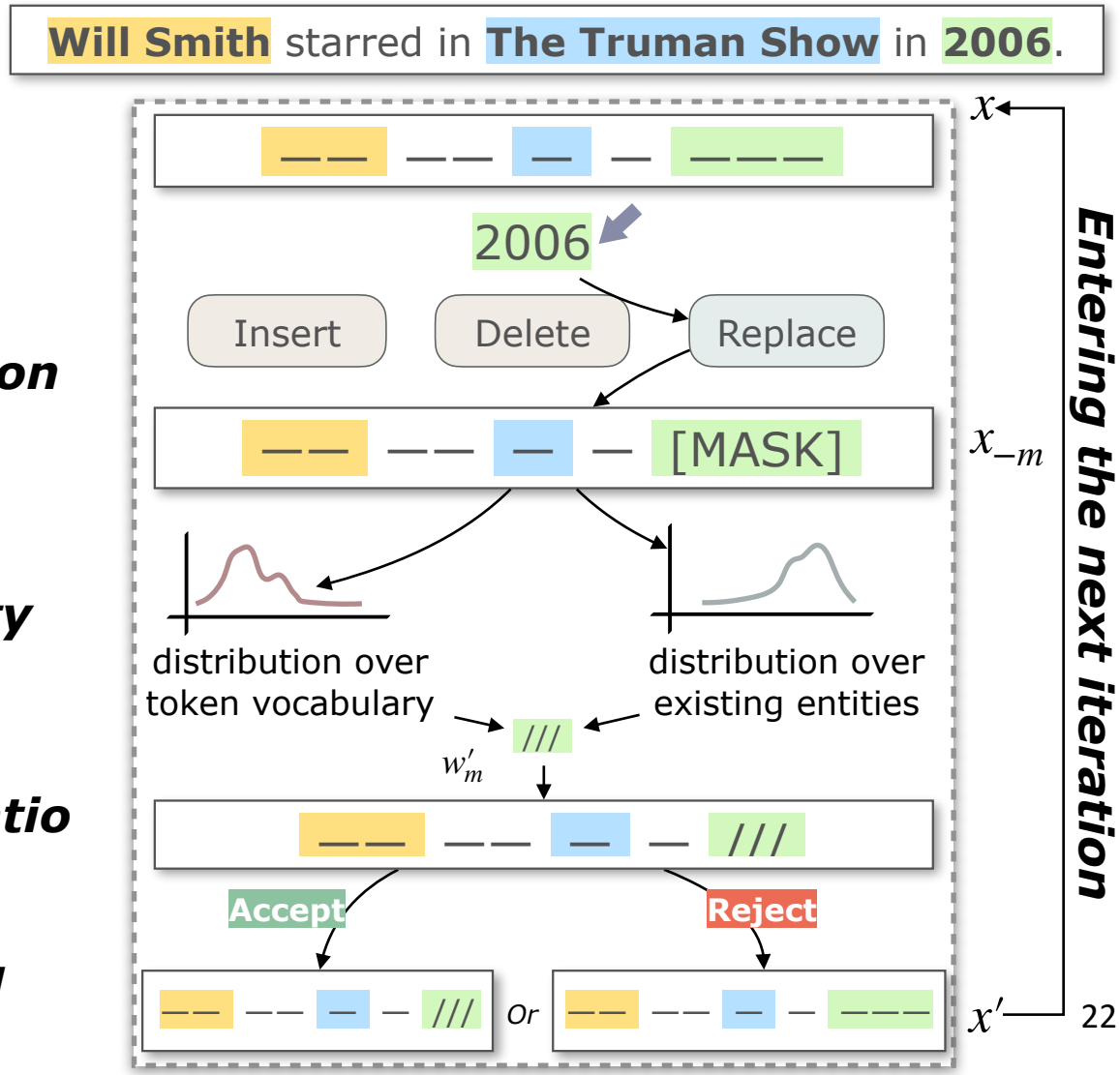
Sample a token/entity

↓

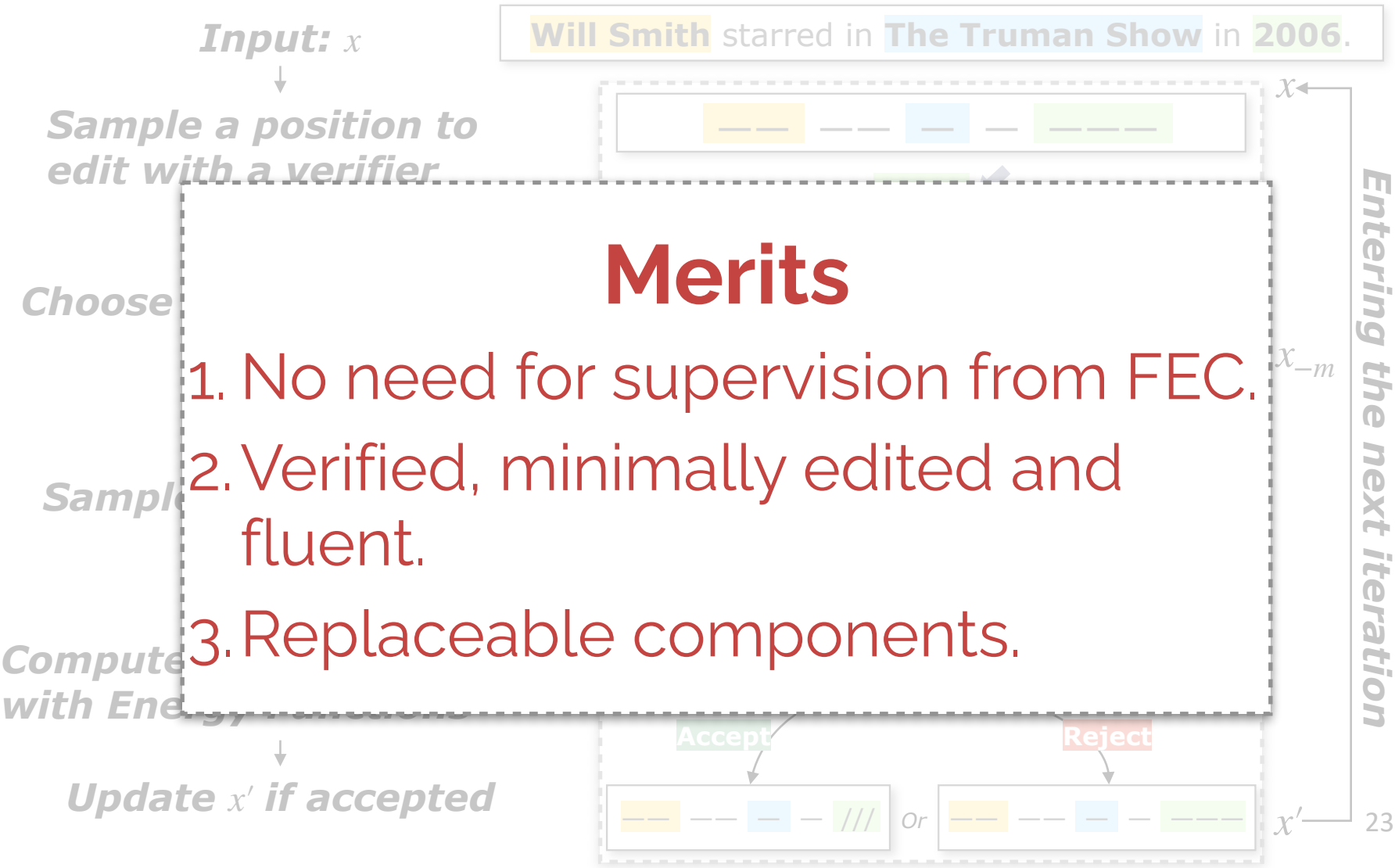
Compute acceptance ratio with Energy Functions

↓

Update x' ***if accepted***



VENCE



Merits

1. No need for supervision from FEC.
2. Verified, minimally edited and fluent.
3. Replaceable components.

Case Study: Editing History

# Iter.	Proposed Claims	Acc.
0	One True Thing is a German film.	-
1	One True Thing is a film.	✓
2	One True Thing is film.	✗
3	One True Thing is a drama film.	✓
4	One True Thing is drama film.	✗
5	One True Thing is American drama film.	✓
6	One True Thing is an American drama film.	✓
7	One True Thing is an American film.	✓
8	One True Thing is an American drama.	✗
9	One True Thing is an 1998 American film.	✗
10	One True Thing is an American.	✗
11	**One True Thing is an American drama film.**	✓
12	One True Thing is an American drama.	✗
13	One True Thing is an American film.	✗
14	One True Thing is a American drama film.	✗
15	One True Thing is an American film.	✗
-	One True Thing is an American film.	Gold

Experiments: Setup

- **Datasets**
  - FECData for FEC (Thorne et al. 2021)
  - FEVER for FV (Thorne et al. 2018)

Label	# Train	# Valid	# Test
SUPPORTED	37,961	1,477	1,593
REFUTED	20,075	2,091	2,289

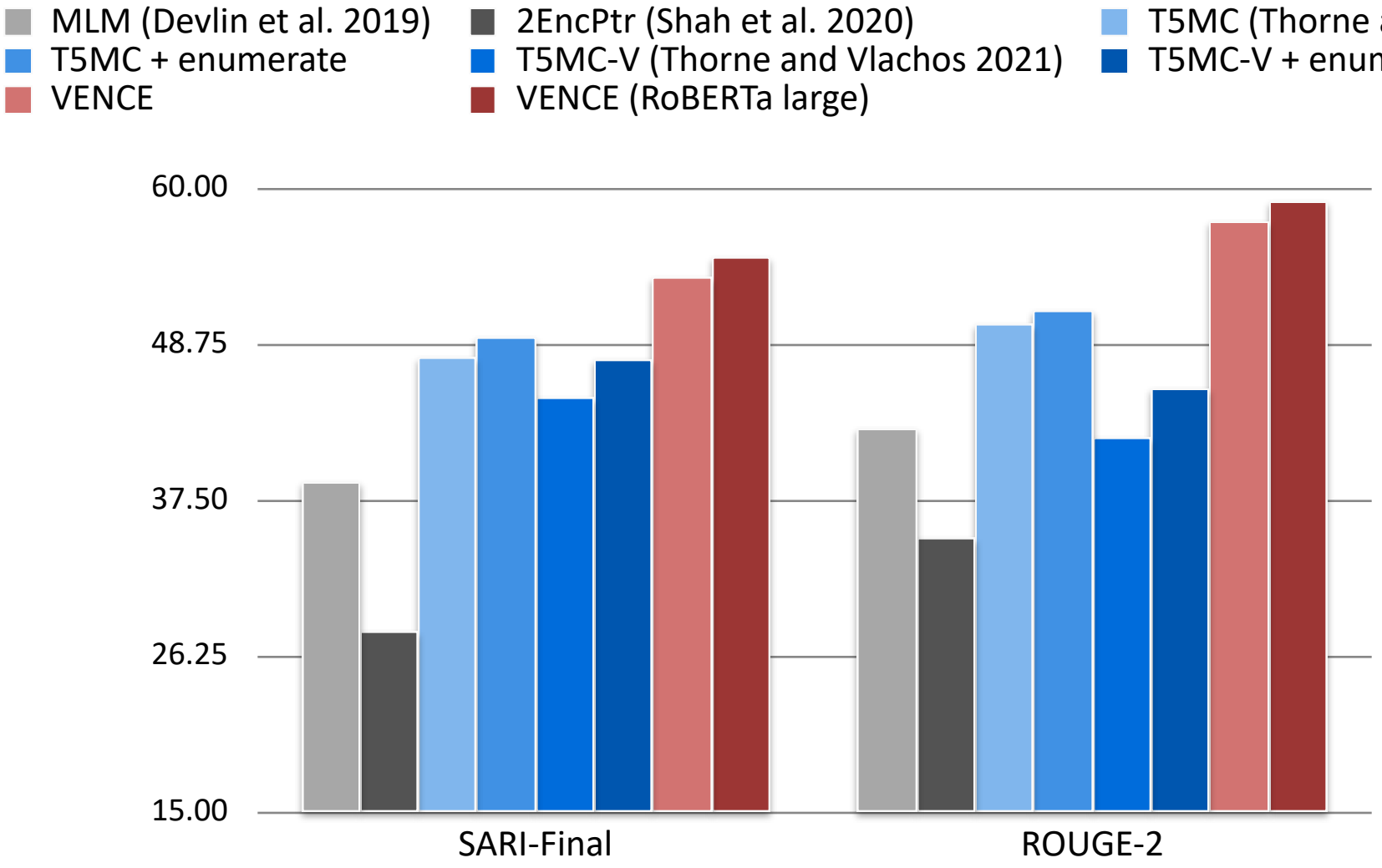
Table 1: Statistics of FECDATA (Thorne and Vlachos 2021), with data sample counts of each split and label.

- **Metrics**
  - SARI scores: the F1 of words being *added*/*deleted*/*kept*
  - Human evaluation (accuracy)

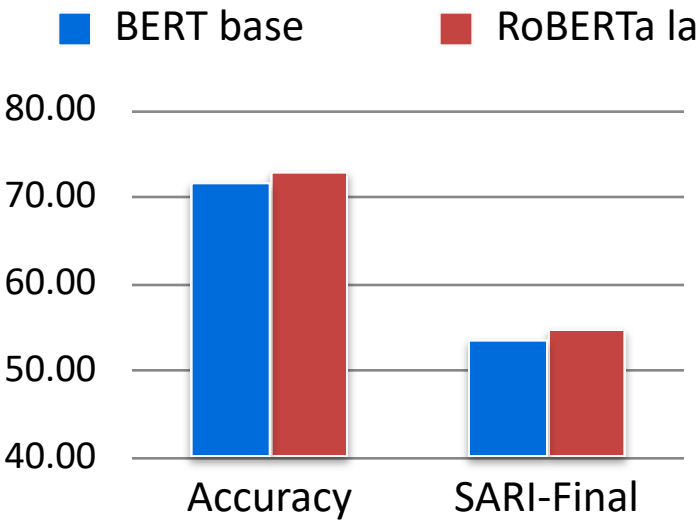
Experiments: Baselines

- **Supervised baselines**
  - T5 (Raffel et al. 2020)
  - EdiT5 (Mallinson et al. 2022)
- ***Distantly-supervised baselines***
  - **DS-1**: Train to propose with evidence-based mask-prediction.
    - *MLM* (Devlin et al. 2019);
    - *2EntPtr* (Shah et al. 2020);
    - ***T5MC*** (Thorne et al. 2021) (+enumerate)
  - **DS-2**: Give a verifier (external discriminative models), e.g., NLI, FV.
    - ***T5MC-V*** (Thorne et al. 2021) (+enumerate)

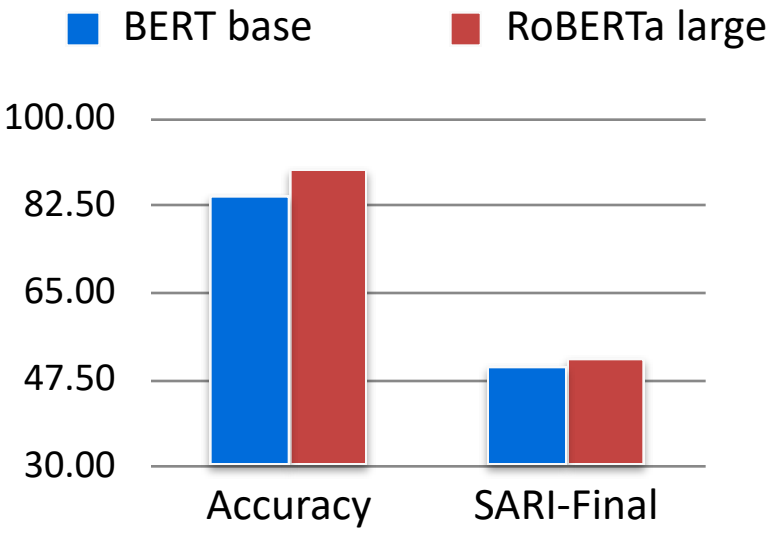
Significant Improvements over Previous DS SoTA



Does better verification leads to better correction?

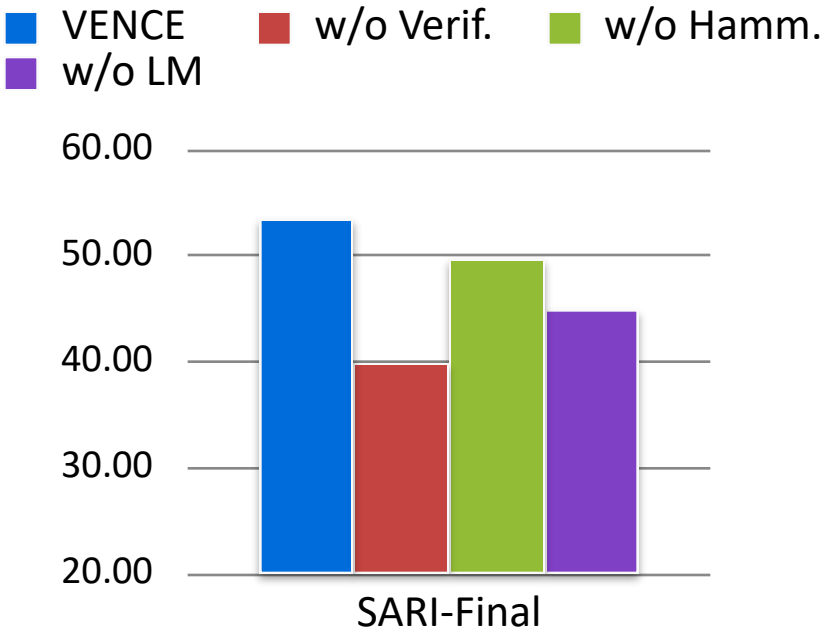


- ID Results on FEVER

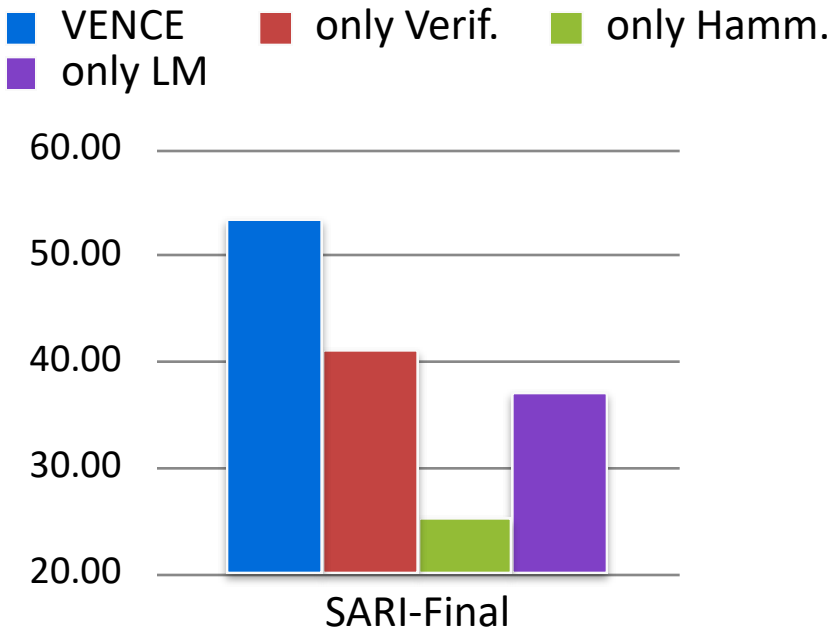


- OOD Results on MultiNLI

Which score contributes the most?

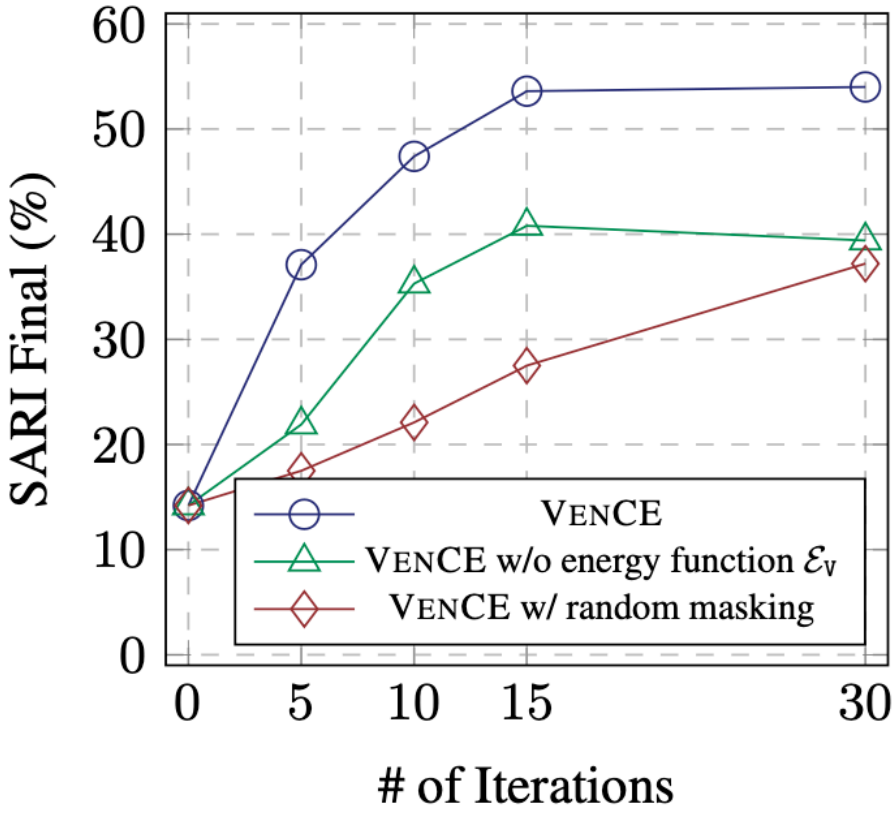


- Leave-one-out



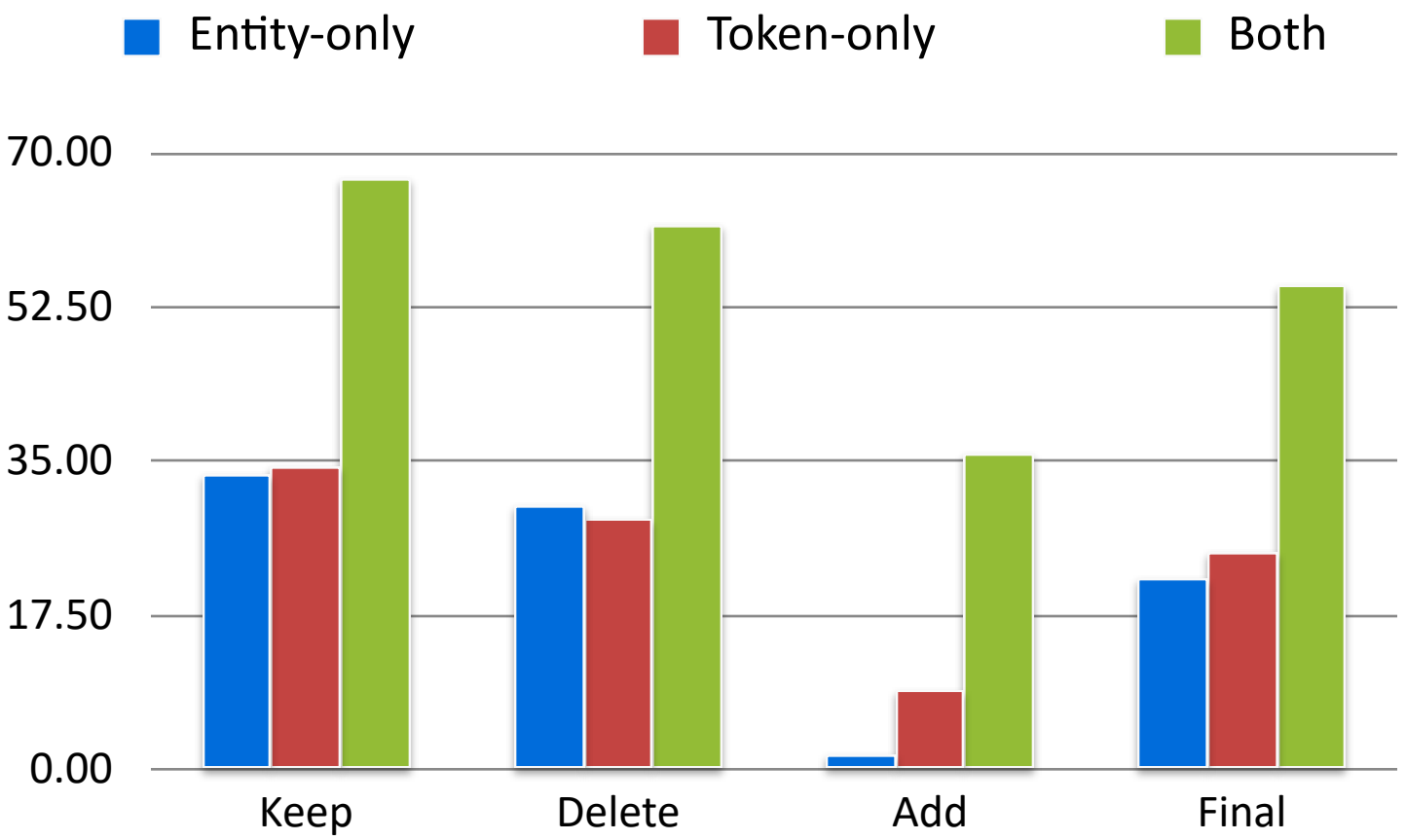
- Only-keep-one

Will more editing iterations help correction?

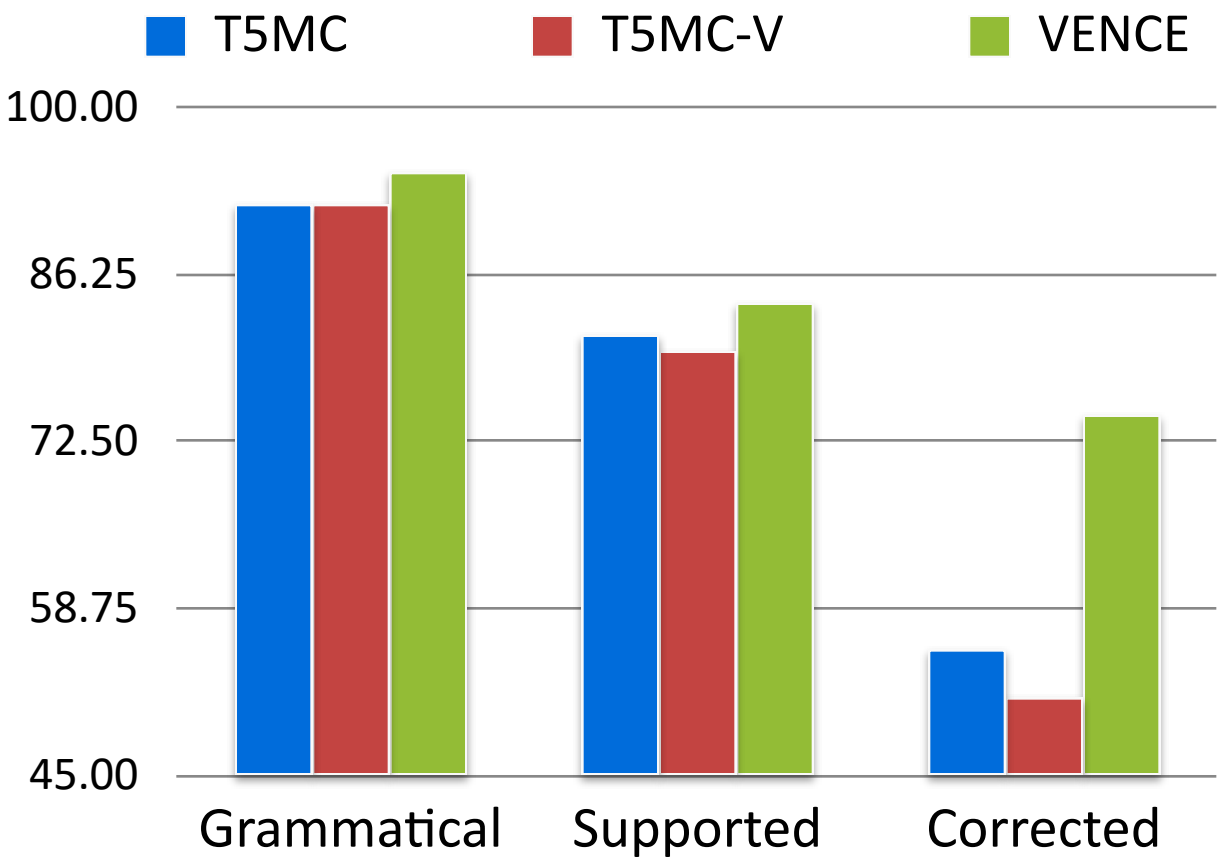


- VENCE converges at Iter #15.
- Performance drop when losing $\mathscr{E}_V(x)$ (fact verification).
- Gradient-based sampling accelerates convergence.

What if we only edit tokens or entities?



Human Evaluation: VENCE productions are not only supported, but with errors corrected.



1. *Is it **grammatically** correct?*
2. *Is it **supported** by evidence?*
3. *Are the original errors **corrected**?*

MLM/T5 may over-erase

Aphrodite is unmarried.

Evidence 1: [Aphrodite] and had many lovers — both gods, such as Ares, and men, such as Anchises. She played a role
in the Eros and Psyche legend, and was both lover and surrogate mother of Adonis. Many lesser beings were said to be children of Aphrodite.

Evidence 2: [Aphrodite] claimed to be her place of birth. In Greek myth, the other gods feared that Aphrodite's beauty might lead to conflict and war, through rivalry for her favours; so Zeus married her off to Hephaestus. Despite this, Aphrodite followed her own inclinations.

T5MC	T5MC-V	VENCE
Aphrodite is both lover and surrogate mother of Adonis.	Aphrodite had many lovers.	Aphrodite is married.
True but semantics change.		**Corrected**

Summary and Takeaways

- Formulate the factual error correction as a **sampling problem**
- VENCE: **Iterative editing (MCMC)** with self-supervised proposals
- Using external fact verification model (LoREN AAAI 2022) as the guidance
- **Limitations and future work**
  - different degrees of factual errors.
  - Larger comprehensive datasets for the FEC task.

Converge to the Truth: Factual Error Correction via Iterative Constrained Editing

Jiangjie Chen, Rui Xu, Wenxuan Zeng,
Changzhi Sun, Lei Li, Yanghua Xiao

For automatic fact verification:
LOREN: Chen et al. AAAI 2022

jjchen19@fudan.edu.cn

https://jiangjiechen.github.io/publication/vence/