UMR IRISA

Institut de Recherche en Informatique
et Systèmes Aléatoires

# Multimodal Misinformation Detection: Overcoming the Training Data Collection Challenge Through Data Generation

**Antoine Chaffin**

Supervisors: Ewa Kijak, Vincent Claveau

Jury: Benoit Favre, Olivier Ferret, Damien Lolive, Claire Gardent, Sylvain Lamprier, Vivien Chappelier

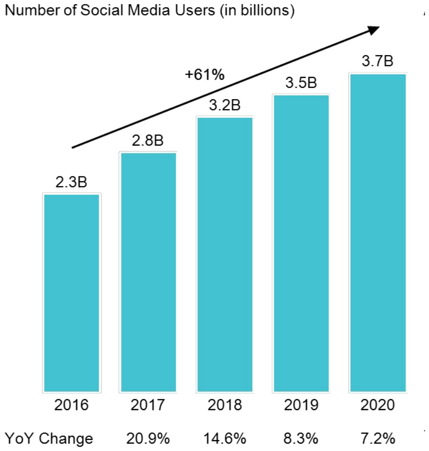
CentraleSupélec · cnrs · ENS rennes · IMT Atlantique · Inria · INSA Rennes · ubs · Université de Rennes 1

IMATAG

# Introduction

# Social networks and multimodal misinformation

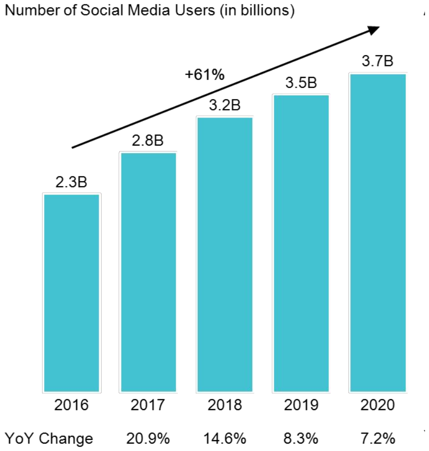
- Growth of social networks usage, used as sources of information
- Every user is a content creator, false information appear



https://www.mekkographics.com/social-media-usage-growth/

# Social networks and multimodal misinformation

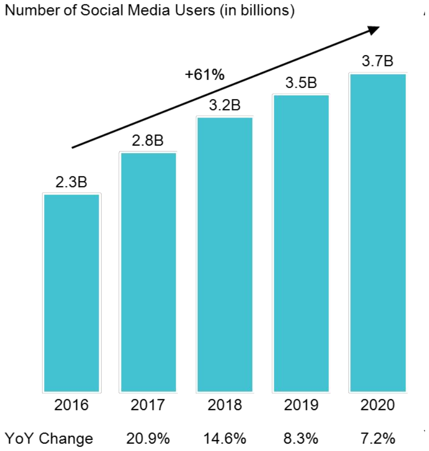
- Growth of social networks usage, used as sources of information
- Every user is a content creator, false information appear
- Numerous and heavy negative consequences

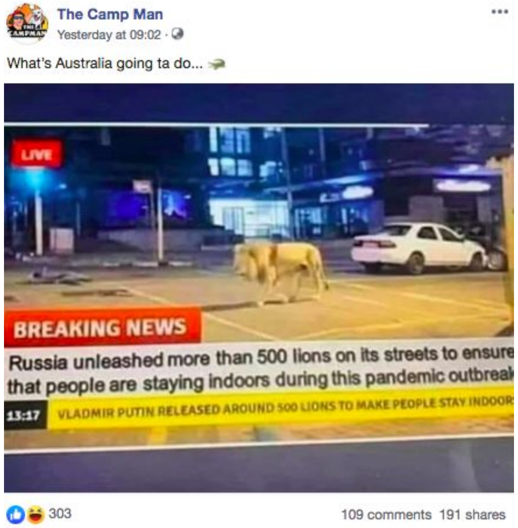


https://www.mekkographics.com/social-media-usage-growth/
https://edition.cnn.com/2020/10/25/world/trump-fake-news-legacy-intl/index.html

# Social networks and multimodal misinformation

- Growth of social networks usage, used as sources of information
- Every user is a content creator, false information appear
- Numerous and heavy negative consequences
- Images coupled with texts to increase traffic and belief
  - Vector of misinformation propagation




The Camp Man
Yesterday at 09:02 ·

What's Australia going ta do...

LIVE

BREAKING NEWS

Russia unleashed more than 500 lions on its streets to ensure that people are staying indoors during this pandemic outbreak

13:17 VLADMIR PUTIN RELEASED AROUND 500 LIONS TO MAKE PEOPLE STAY INDOOR

303
109 comments 191 shares

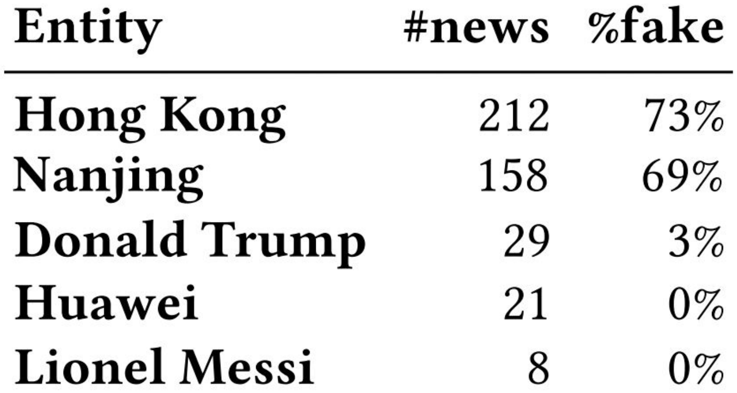
https://www.mekkographics.com/social-media-usage-growth/
https://edition.cnn.com/2020/10/25/world/trump-fake-news-legacy-intl/index.html
https://factcheck.afp.com/photo-was-taken-south-africa-2016-it-unrelated-coronavirus-pandemic

# Suitable dataset

- Manual fact-checking too long and costly, automatic detection needed
- Pre-trained models **require fine-tuning to achieve best performance**
- High-capacity models prone to overfitting (esp. multimodal)

# Suitable dataset

- Manual fact-checking too long and costly, automatic detection needed
- Pre-trained models **require fine-tuning to achieve best performance**
- High-capacity models prone to overfitting (esp. multimodal)
- Biased datasets yield **non-generalizable features**
  - Highly variable performances



| Entity | #news | %fake |
|---|---|---|
| Hong Kong | 212 | 73% |
| Nanjing | 158 | 69% |
| Donald Trump | 29 | 3% |
| Huawei | 21 | 0% |
| Lionel Messi | 8 | 0% |

Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, Fuzhen Zhuang. “Generalizing to the Future: Mitigating Entity Bias in Fake News Detection”. 2022

# Language modeling

- Reproduce example sequences
- Probability of the **next word given past ones**

“The cat drinks milk”

The → cat

The cat → drinks

The cat drinks → milk

Teacher forcing[1]

[1] Ronald J. Williams, David Zipser. “A Learning Algorithm for Continually Running Fully Recurrent Neural Networks”. 1989

# Language modeling

- Reproduce example sequences
- Probability of the **next word given past ones**
- **Iteratively add tokens** to produce text

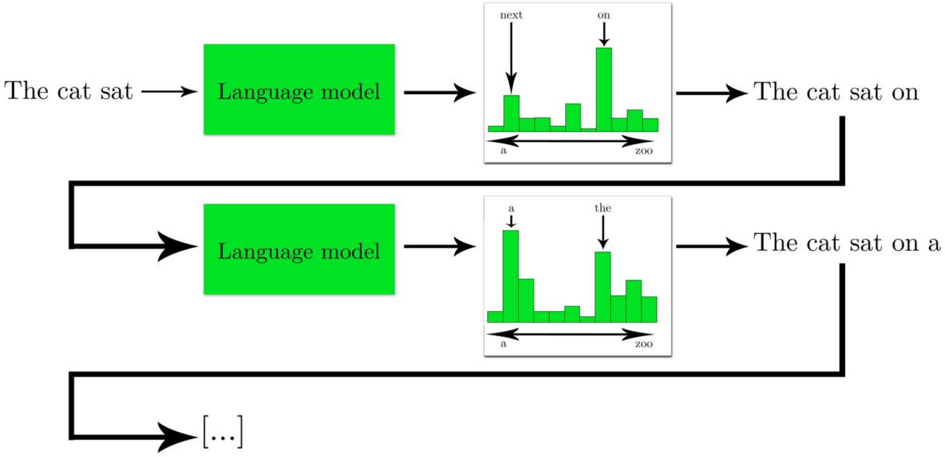
“The cat drinks milk”

The → cat

The cat → drinks

The cat drinks → milk

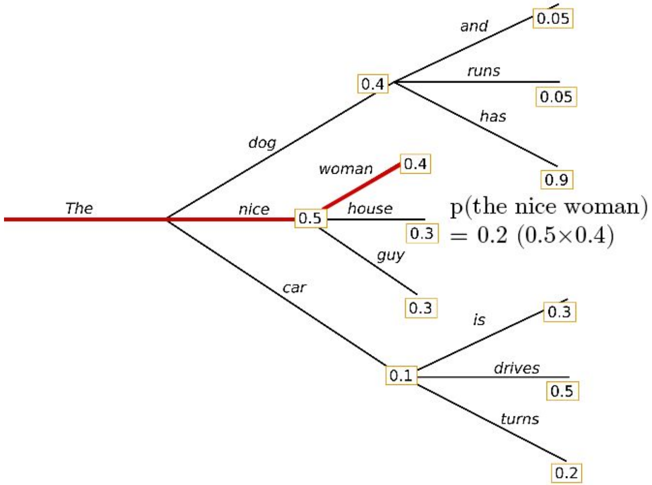
Teacher forcing[1]



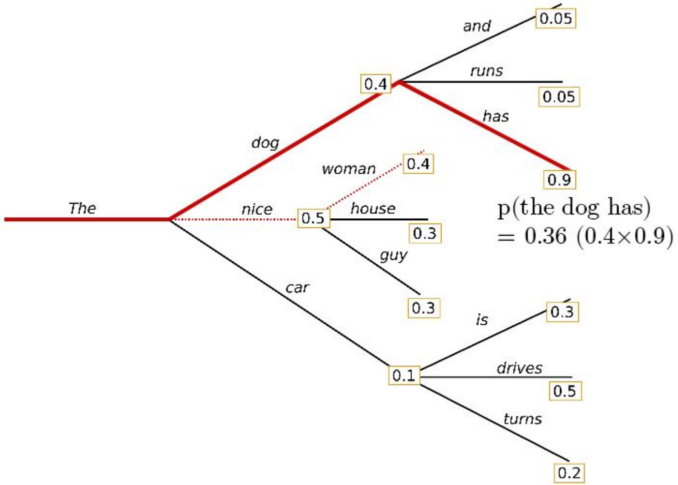
[1] Ronald J. Williams, David Zipser. “A Learning Algorithm for Continually Running Fully Recurrent Neural Networks”. 1989

# Deterministic decoding

- Text generation can be seen as **tree exploration**
- **Root is the prompt**, each node corresponds to its **parent sequence with an additional token**
- Search highest likelihood sequence



Greedy search



Beam search

https://huggingface.co/blog/how-to-generate

# Generating training data

**Pakistan police clash with doctors protesting lack of virus protection gear**

(Reuters) - Riot police wielding batons used force to break up a protest by Pakistani doctors and medical staff against a lack of gear to protect against coronavirus, arresting dozens of medics who say the government has failed to deliver promised supplies. Reuters journalists at the scene, in the southwestern city of Quetta, saw hundreds of doctors and paramedics, some in face masks and scrubs, chanting their demands. Some were dragged off by riot police in helmets, armed with rifles and batons. A senior police official said 30 protesters had been arrested for defying a ban on public gatherings imposed during a lockdown to fight the spread of the virus.

[...]

**Entity recognition & swapping**

**Indian police clash with doctors protesting lack of virus protection gear**

**Text generation**

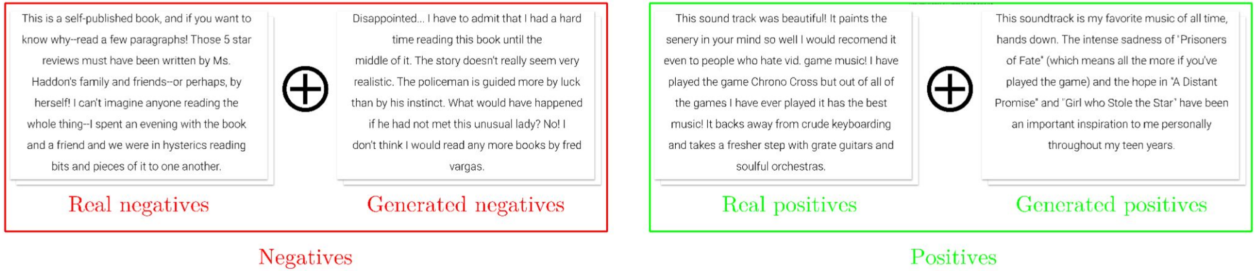
(Reuters) - Indian doctors protested in front of Parliament on Friday in a rare show of unity against a government that they say is not sufficiently protecting them against the country’s dreaded AIDS virus. India’s healthcare system is struggling to cope with HIV/AIDS and is failing to achieve national goals for both treatment and prevention, one recent study showed. As the healthcare system struggles, doctors are fighting to boost testing of the BRCA1 and BRCA2 genes to find those carriers of the HIV virus who can take anti-retroviral drugs. In the first major test of health financing, about half the 13,000 doctors in Kolkata went on a strike in March this year to protest at not being given protective uniforms and rubber gloves during AIDS tests.

[...]

https://www.reuters.com/article/us-health-coronavirus-pakistan-healthcar/pakistan-police-clash-with-doctors-protesting-lack-of-virus-protection-gear-idUSKBN2101HJ

# Synthetic textual data augmentation

- Generated texts can be used for data augmentation★



This is a self-published book, and if you want to know why--read a few paragraphs! Those 5 star reviews must have been written by Ms. Haddon's family and friends--or perhaps, by herself! I can't imagine anyone reading the whole thing--I spent an evening with the book and a friend and we were in hysterics reading bits and pieces of it to one another.

Disappointed... I have to admit that I had a hard time reading this book until the middle of it. The story doesn't really seem very realistic. The policeman is guided more by luck than by his instinct. What would have happened if he had not met this unusual lady? No! I don't think I would read any more books by fred vargas.

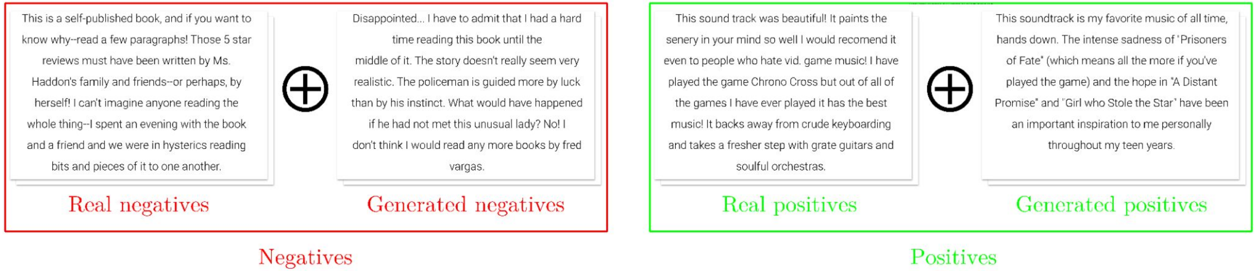
This sound track was beautiful! It paints the senery in your mind so well I would recomend it even to people who hate vid. game music! I have played the game Chrono Cross but out of all of the games I have ever played it has the best music! It backs away from crude keyboarding and takes a fresher step with grate guitars and soulful orchestras.

This soundtrack is my favorite music of all time, hands down. The intense sadness of 'Prisoners of Fate" (which means all the more if you've played the game) and the hope in "A Distant Promise" and 'Girl who Stole the Star" have been an important inspiration to me personally throughout my teen years.

★ Vincent Claveau, **Antoine Chaffin**, Ewa Kijak. "Generating artificial texts as substitution or complement of training data". *LREC*. 2022

# Synthetic textual data augmentation

- Generated texts can be used for data augmentation★



This is a self-published book, and if you want to know why--read a few paragraphs! Those 5 star reviews must have been written by Ms. Haddon's family and friends--or perhaps, by herself! I can't imagine anyone reading the whole thing--I spent an evening with the book and a friend and we were in hysterics reading bits and pieces of it to one another.

Real negatives

⊕

Disappointed... I have to admit that I had a hard time reading this book until the middle of it. The story doesn't really seem very realistic. The policeman is guided more by luck than by his instinct. What would have happened if he had not met this unusual lady? No! I don't think I would read any more books by fred vargas.

Generated negatives

Negatives

This sound track was beautiful! It paints the senery in your mind so well I would recomend it even to people who hate vid. game music! I have played the game Chrono Cross but out of all of the games I have ever played it has the best music! It backs away from crude keyboarding and takes a fresher step with grate guitars and soulful orchestras.

Real positives

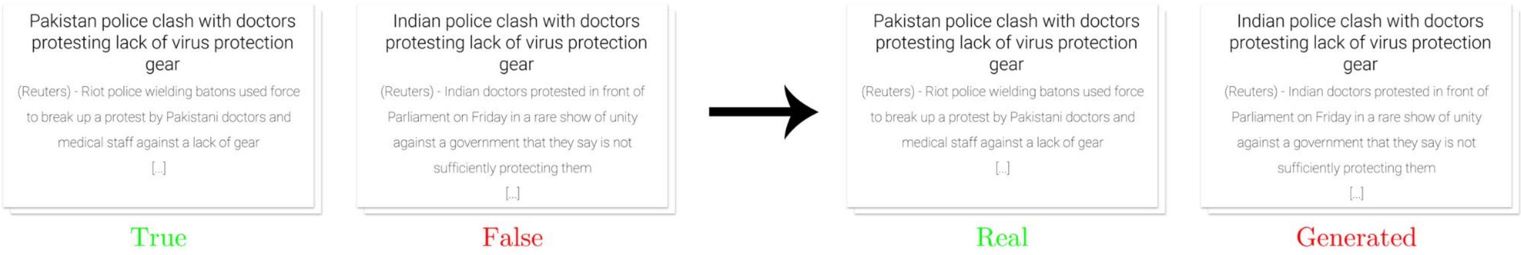
⊕

This soundtrack is my favorite music of all time, hands down. The intense sadness of 'Prisoners of Fate" (which means all the more if you've played the game) and the hope in "A Distant Promise" and "Girl who Stole the Star" have been an important inspiration to me personally throughout my teen years.

Generated positives

Positives

- Class **"false"** only composed of generated data, **"true"** only composed of real articles
  - Classes become **"real"** and **"generated"**



Pakistan police clash with doctors protesting lack of virus protection gear

(Reuters) - Riot police wielding batons used force to break up a protest by Pakistani doctors and medical staff against a lack of gear

[...]

True

Indian police clash with doctors protesting lack of virus protection gear

(Reuters) - Indian doctors protested in front of Parliament on Friday in a rare show of unity against a government that they say is not sufficiently protecting them

[...]

False

→

Pakistan police clash with doctors protesting lack of virus protection gear

(Reuters) - Riot police wielding batons used force to break up a protest by Pakistani doctors and medical staff against a lack of gear

[...]

Real

Indian police clash with doctors protesting lack of virus protection gear

(Reuters) - Indian doctors protested in front of Parliament on Friday in a rare show of unity against a government that they say is not sufficiently protecting them

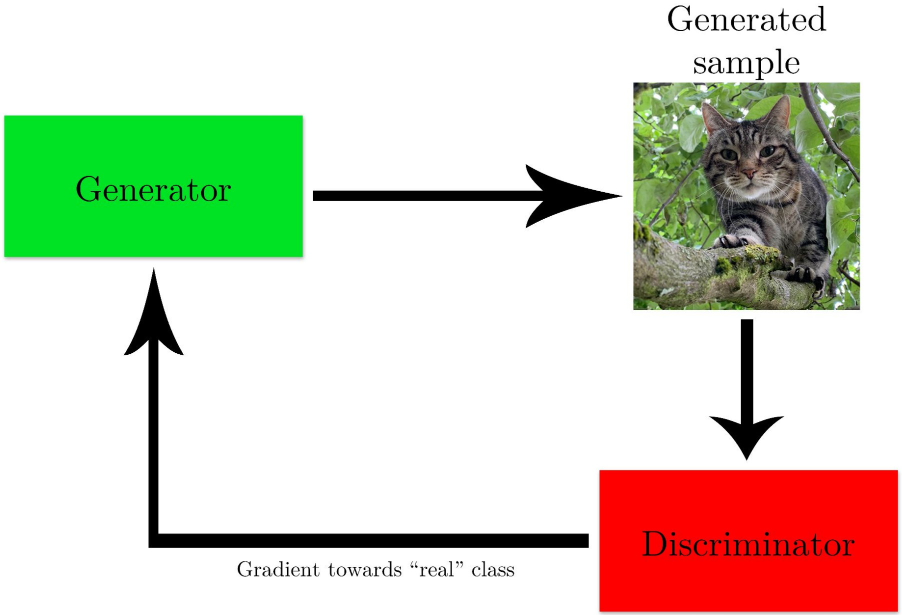
[...]

Generated

★ Vincent Claveau, **Antoine Chaffin**, Ewa Kijak. "Generating artificial texts as substitution or complement of training data". *LREC*. 2022

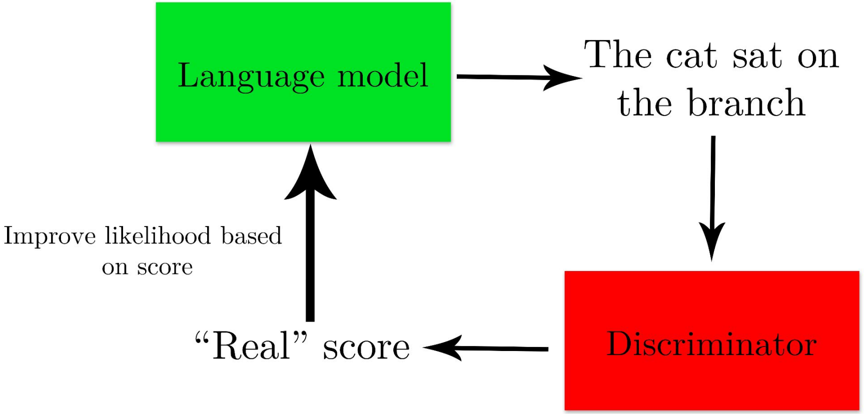
# Generative Adversarial Network

- Discriminator can detect generated data
- Use this information to **guide the generator towards undetectable samples**
- Both networks improve together



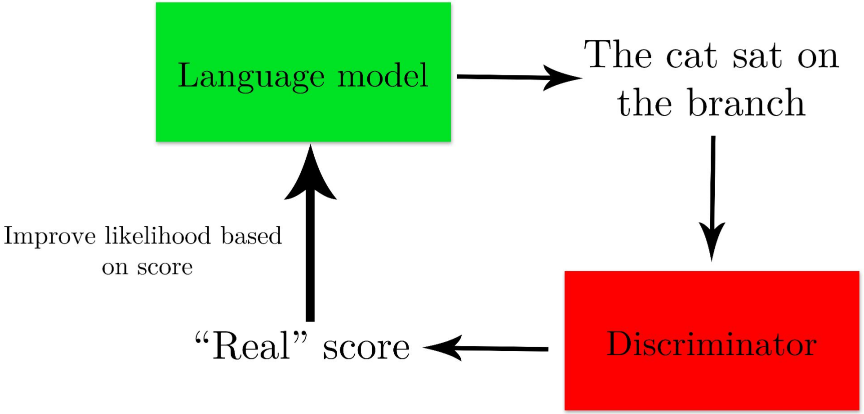
# Textual Generative Adversarial Network

- Discretization prevents the direct backpropagation of the gradient
- ⇒ Reinforcement learning with discriminator scores as rewards



# Textual Generative Adversarial Network

- Discretization prevents the direct backpropagation of the gradient
- ⇒ Reinforcement learning with discriminator scores as rewards
- Improve likelihood of highest-scoring sequences
  - Sparse rewards

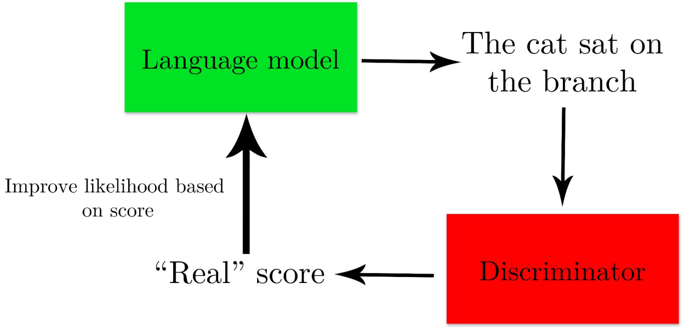


$$\nabla_\theta L_\theta \underbrace{(x)}_{\text{Sample from the generator}} = - \overbrace{r(x)}^{\text{Reward}} \nabla_\theta \overbrace{\log p_\theta (x)}^{\text{Likelihood}}$$

# Cooperative textual generation

- Information from the discriminator:

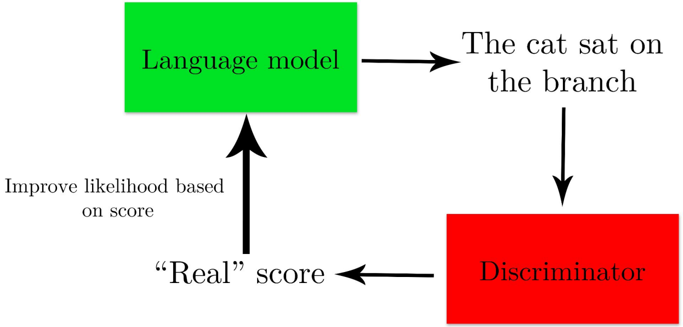
- Can be used **to train** the language model (**adversarial** approach)



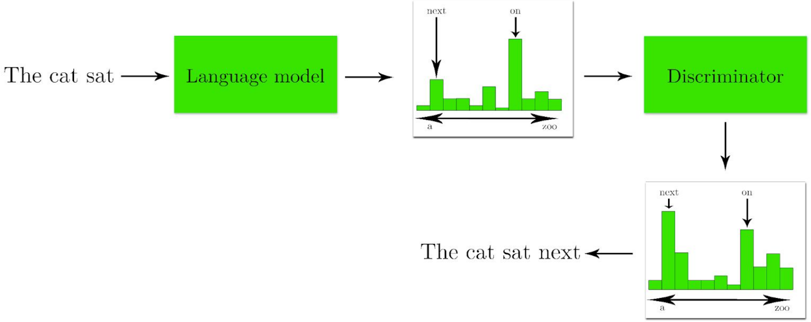
# Cooperative textual generation

- Information from the discriminator:

- Can be used **to train** the language model (**adversarial** approach)



- But also to **guide the decoding (cooperative** approach)



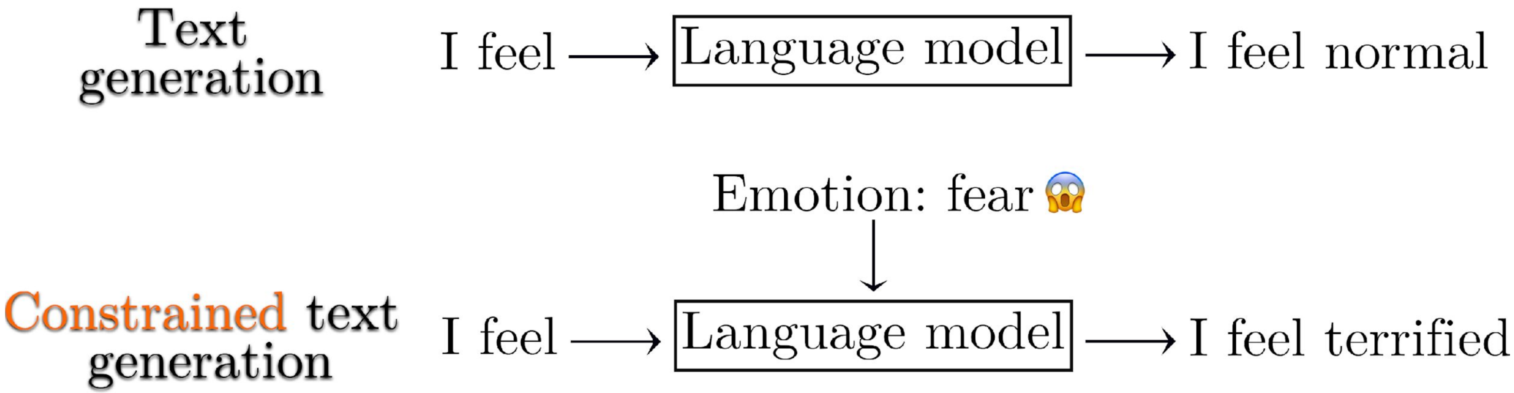
- Higher (denser) rewards

1. PPL-MCTS: Constrained Textual Generation Through Discriminator-Guided MCTS Decoding
2. Generative Cooperative Networks for Natural Language Generation
3. Distinctive Image Captioning: Leveraging Ground Truth Captions in CLIP Guided Reinforcement Learning
4. Conclusion & perspectives

1. PPL-MCTS: Constrained Textual Generation Through Discriminator-Guided MCTS Decoding
2. Generative Cooperative Networks for Natural Language Generation
3. Distinctive Image Captioning: Leveraging Ground Truth Captions in CLIP Guided Reinforcement Learning
4. Conclusion & perspectives

# Constrained textual generation

- Few options to control the generation besides the **prompt**
- Adding some **constraints** is useful to control various aspects (writing style, emotion/polarity, detoxification, etc.)



Text generation

I feel → Language model → I feel normal

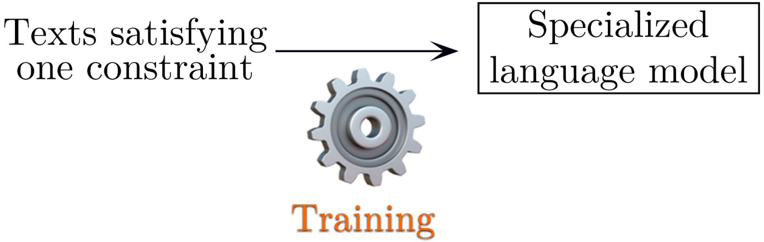
Emotion: fear 😱

Constrained text generation

I feel → Language model → I feel terrified

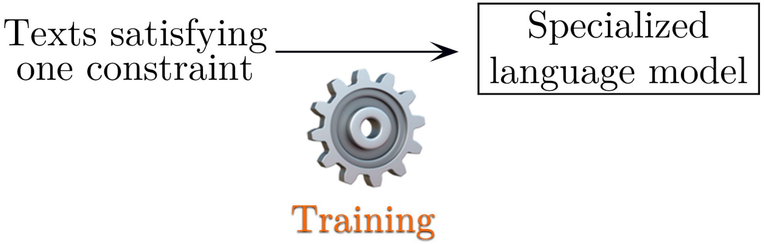
# Previous work

- Language models (LMs) tuning
  - Train and store **one model for each constraint**
  - **Very costly** when even possible for very large LM



# Previous work

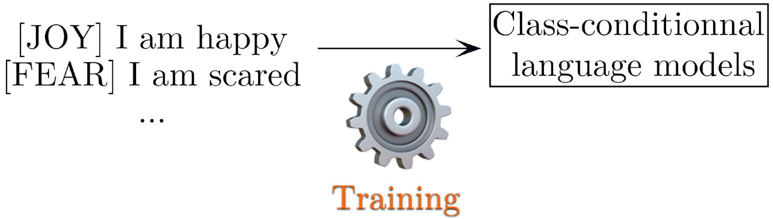
- Language models (LMs) tuning
  - Train and store **one model for each constraint**
  - **Very costly** when even possible for very large LM
- Class-conditional language models
  - Add a **control code** before texts
  - **Training/tuning for any new additional constraint**



Texts satisfying one constraint

Specialized language model

Training



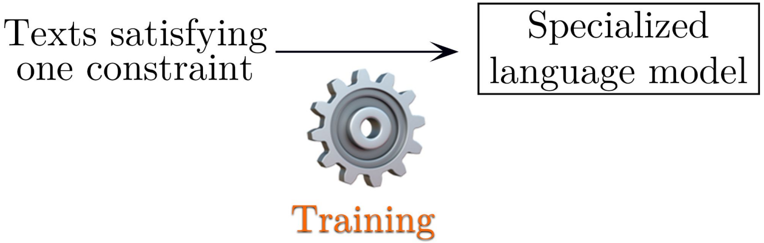
[JOY] I am happy
[FEAR] I am scared
...

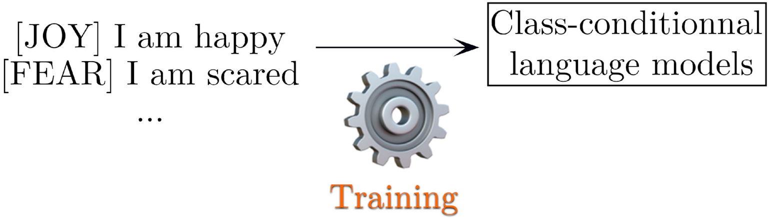
Class-conditionnal language models

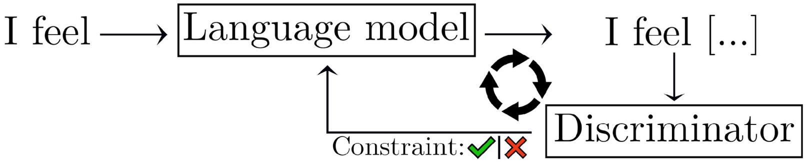
Training

# Previous work

- Language models (LMs) tuning
  - Train and store **one model for each constraint**
  - **Very costly** when even possible for very large LM
- Class-conditional language models
  - Add a **control code** before texts
  - **Training/tuning for any new additional constraint**
- Cooperative generation
  - Change the LM distribution based on **scores from a discriminator**



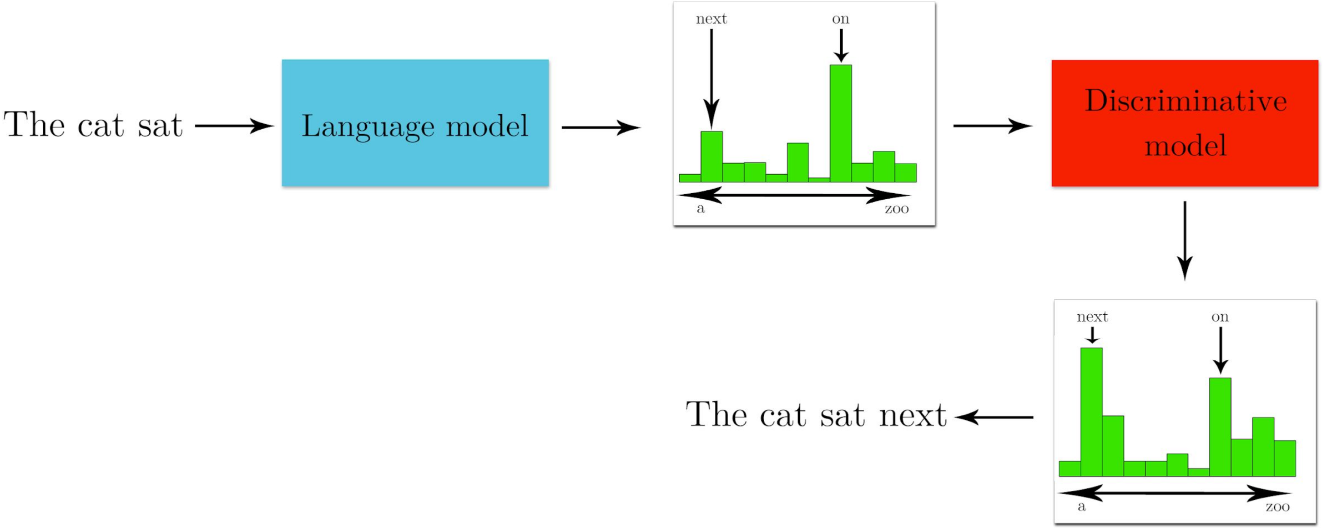




# Cooperative textual generation

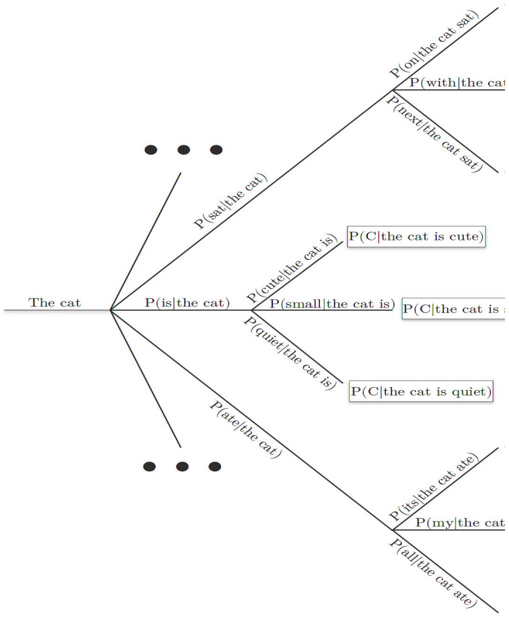
- Guide the generation using the **score of an external model**
- Generate text following the conditional distribution (product of the language model likelihood and the score of the discriminative model)

$$p(x \mid c) \propto p(x) * p(c \mid x)$$



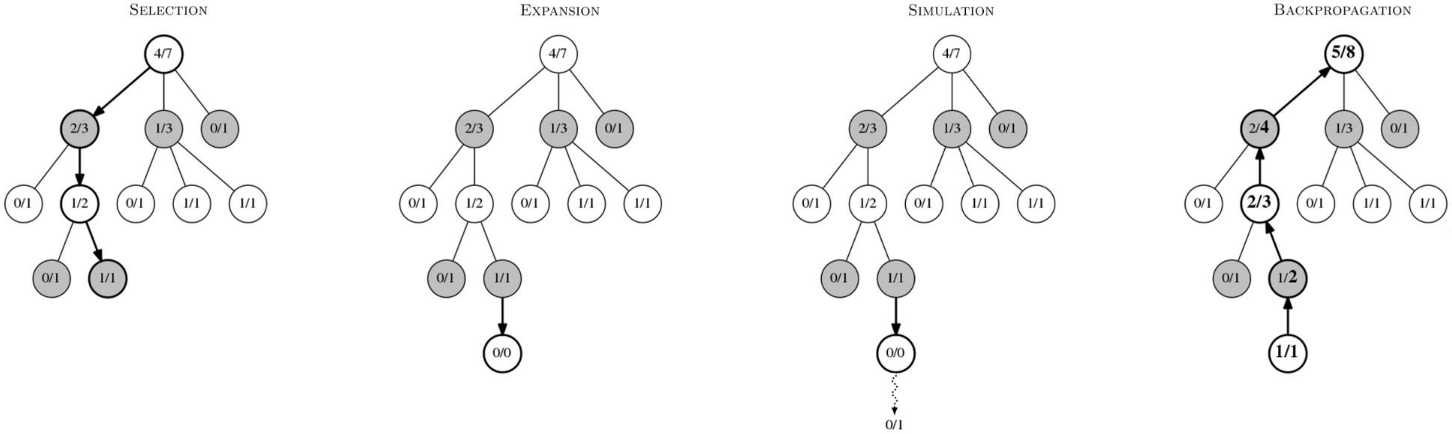
# Short term decisions vs. long-term result

- Generation is **iterative**, one token is produced after the other
  - Previous studies focus on the next token to emit (lack of **long-term vision**)
- Words' meaning are **context-dependent**
  - This book is **awfully**
    - Good/bad ?
  - This made me **cry**
    - Joy/Sadness ?
- **Short-term** decisions might not be optimal in the **long run**



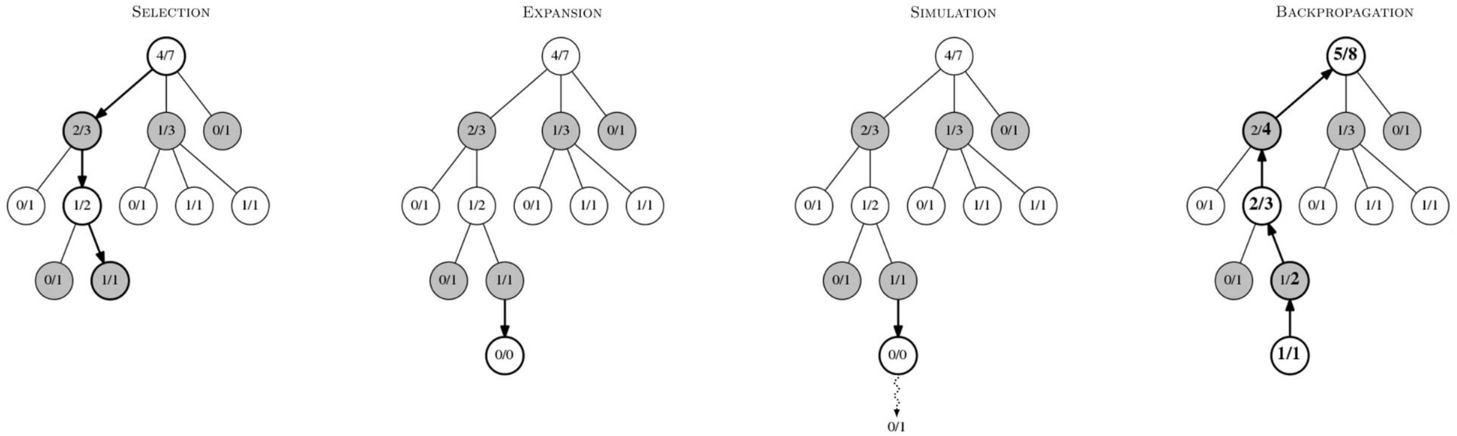
# Monte Carlo Tree Search

- Heuristic-based iterative algorithm that uses randomness to solve deterministic problems when the **search space is too large**



# Monte Carlo Tree Search

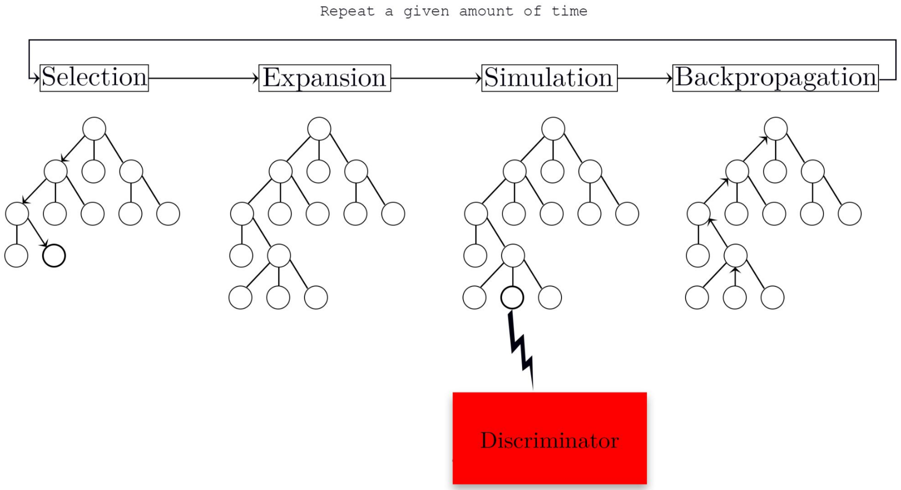
- Heuristic-based iterative algorithm that uses randomness to solve deterministic problems when the **search space is too large**



- Compromise between **exploiting good sequences** and **exploring promising ones**
- Score of a node is defined by children’s (simulation)
  - **Short-term decisions to optimize a long-term result**

# PPL-MCTS: Plug and Play Language-Monte Carlo Tree Search

- Monte Carlo Tree Search (MCTS) properties:
  1. **Long-term vision**: scores the next token using finished sequences (rollout)
  2. **Efficient**: exploration of sub-optimal paths has an upper bound
  3. **Modular**: outputs a solution according to the computational budget
  4. **Plug and play**: can be used with any LM and discriminator without any tuning



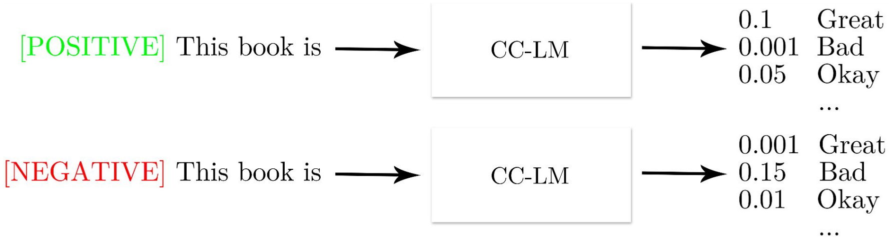
# Datasets

- Two tasks: **polarity** 😍😡 and **emotion** 😠😥😃😱💏
- Two languages: **French** and **English** 🇫🇷🇬🇧

| amazon_polarity | CLS (FLUE) | Emotion |
|---|---|---|
| **[POSITIVE]** Stuning even for the non-gamer. This sound track was beautiful! It paints the senery in your mind so well I would recomend it even to people who hate vid. game music! I have played the game Chrono Cross but out of all of the games I have ever played it has the best music! It backs away from crude keyboarding and takes a fresher step with grate guitars and soulful orchestras. It would impress anyone who cares to listen! ^_^ | **[POSITIVE]** Robert Downey Jr en héros de Marvel? Ca apparait de prime abord complètement improbable et après avoir vu le film on se dit que personne d'autre n'aurait pu jouer le rôle d'Iron Man. En effet le film évite les clichés à la testostérone habituel des supers héros, ajoute une très bonne dose d'humour et de glamour, et propose un scénario original (sauf la fin que je trouve bidon). Au final Iron Man est un très bon film d'action qui parvient à renouveller le genre et où les effets spéciaux supportent très bien l'histoire (et non l'inverse). Bref un très bon moment en perspective à découvrir**!** | **[SADNESS]** ive been feeling a little burdened lately wasnt sure why that was |

# Baselines

- Class-conditional language models **(CC-LMs)**[1]



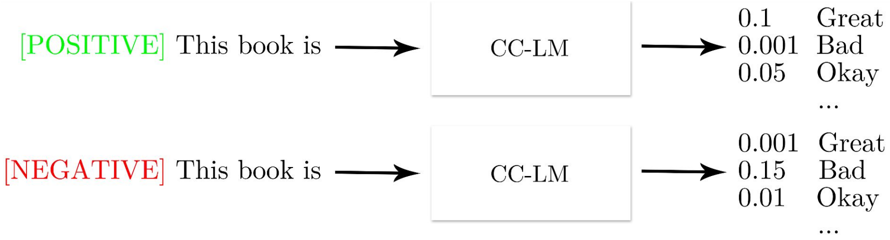
[POSITIVE] This book is → CC-LM →
0.1 Great
0.001 Bad
0.05 Okay
...

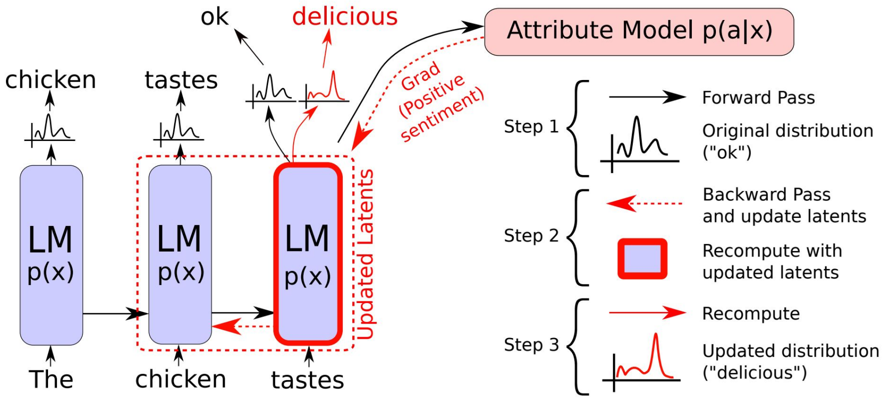
[NEGATIVE] This book is → CC-LM →
0.001 Great
0.15 Bad
0.01 Okay
...

[1] Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, Richard Socher. “CTRL: A Conditional Transformer Language Model for Controllable Generation”. 2019

# Baselines

- Class-conditional language models **(CC-LMs)**[1]
- **PPLM**[2]: use discriminator scores to update LM hidden states

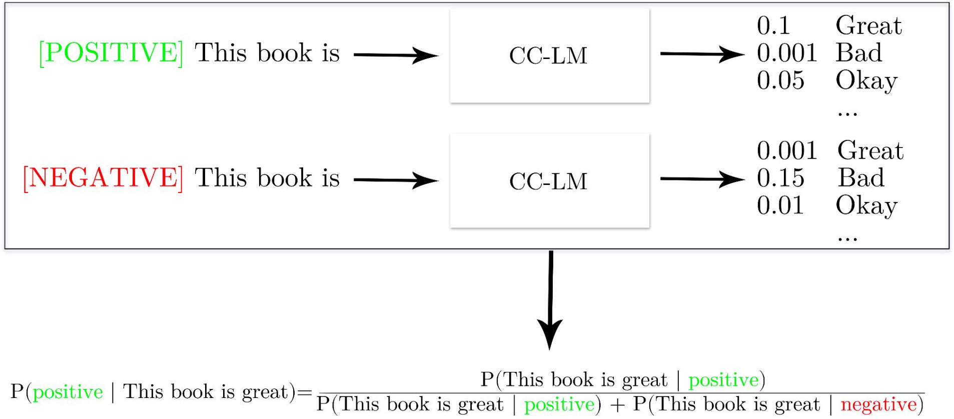



ok
delicious
Attribute Model p(a|x)
chicken
tastes
Grad (Positive sentiment)
LM p(x)
LM p(x)
LM p(x)
Updated Latents
The
chicken
tastes
Step 1
Forward Pass
Original distribution ("ok")
Step 2
Backward Pass and update latents
Recompute with updated latents
Step 3
Recompute
Updated distribution ("delicious")

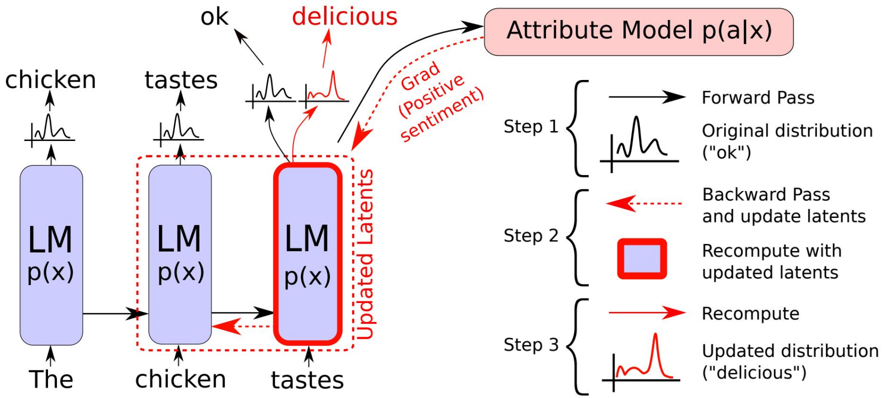
[1] Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, Richard Socher. “CTRL: A Conditional Transformer Language Model for Controllable Generation”. 2019
[2] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, Rosanne Liu. “Plug and Play Language Models: A Simple Approach to Controlled Text Generation”. 2019

# Baselines

- Class-conditional language models **(CC-LMs)**[1]
- **PPLM**[2]: use discriminator scores to update LM hidden states
- Generative Discriminators **(GeDi)**[3]: leverage CC-LMs to get classification scores over the whole vocabulary



$$\mathrm{P}(\text{positive} \mid \text{This book is great}) = \frac{\mathrm{P}(\text{This book is great} \mid \text{positive})}{\mathrm{P}(\text{This book is great} \mid \text{positive}) + \mathrm{P}(\text{This book is great} \mid \text{negative})}$$


ok
delicious
Attribute Model p(a|x)
chicken
tastes
Grad (Positive sentiment)
LM p(x)
LM p(x)
LM p(x)
Updated Latents
The
chicken
tastes
Step 1
Forward Pass
Original distribution ("ok")
Step 2
Backward Pass and update latents
Recompute with updated latents
Step 3
Recompute
Updated distribution ("delicious")

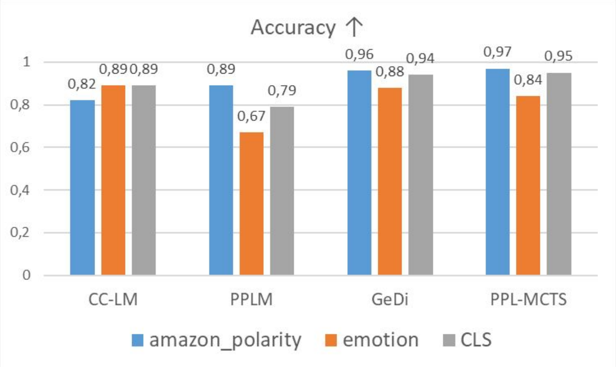
[1] Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, Richard Socher. “CTRL: A Conditional Transformer Language Model for Controllable Generation”. 2019
[2] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, Rosanne Liu. “Plug and Play Language Models: A Simple Approach to Controlled Text Generation”. 2019
[3] Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, Nazneen Fatema Rajani. “GeDi: Generative Discriminator Guided Sequence Generation”. 2020

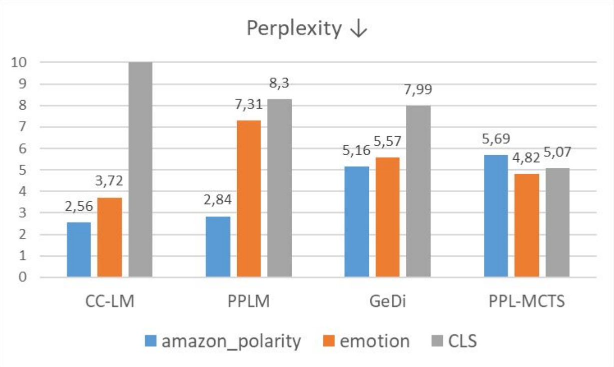
# Results

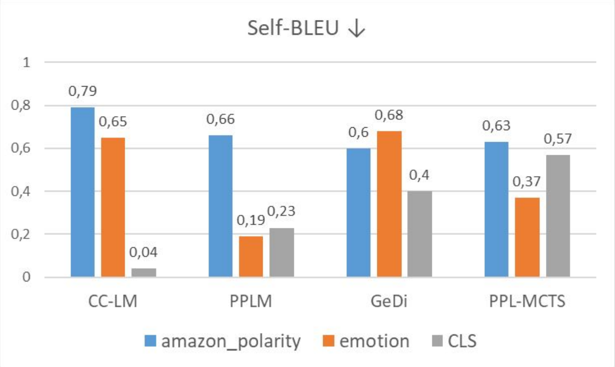
- **Automatic metrics**
  1. **Accuracy**: samples belong to the target class 🎯
  2. **Perplexity**: samples are well written ✍️
  3. **Self-BLEU**: there is enough diversity across samples 📕📗📘

# Results

- **Automatic metrics**
  1. **Accuracy**: samples belong to the target class 🎯
  2. **Perplexity**: samples are well written ✍
  3. **Self-BLEU**: there is enough diversity across samples 📕📗📘
- PPL-MCTS yields **state-of-the-art results on both tasks and languages**
  - Matching GeDi performance that **rely on tuned CC-LMs**
  - Human evaluation **confirmed the results**



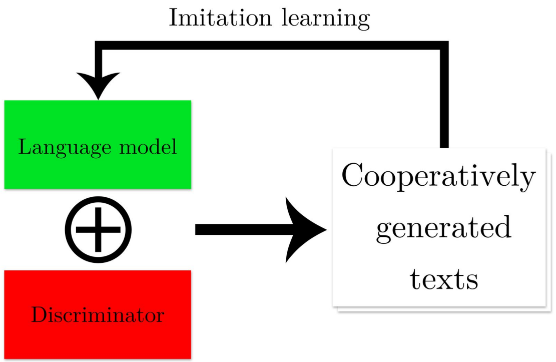




1. PPL-MCTS: Constrained Textual Generation Through Discriminator-Guided MCTS Decoding
2. **Generative Cooperative Networks for Natural Language Generation**
3. Distinctive Image Captioning: Leveraging Ground Truth Captions in CLIP Guided Reinforcement Learning
4. Conclusion & perspectives

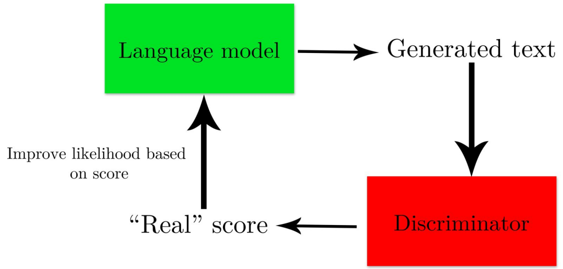
# SelfGAN

- Training via **imitation learning** using cooperatively generated texts



SelfGAN

$$\nabla_\theta L_\theta \underbrace{(x)}_{\text{Cooperative sample}} = -\nabla_\theta \overbrace{\log p_\theta (x)}^{\text{Likelihood}}$$



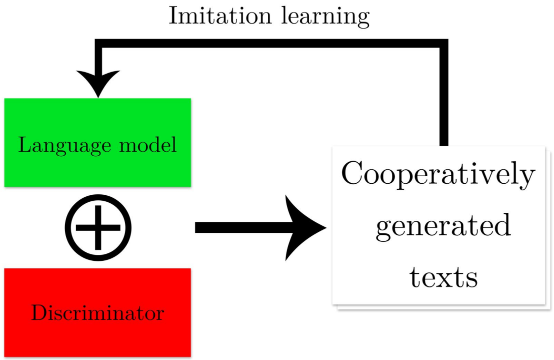
Discrete GAN

$$\nabla_\theta L_\theta \underbrace{(x)}_{\text{Sample from the generator}} = -\overbrace{D(x)}^{\text{Discriminator score}} \nabla_\theta \overbrace{\log p_\theta (x)}^{\text{Likelihood}}$$

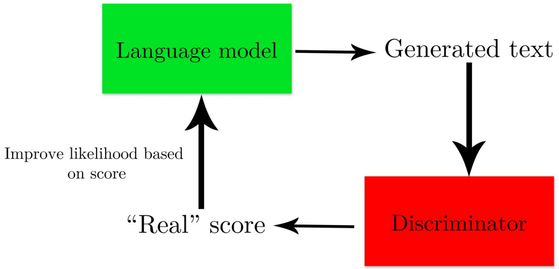
Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano. "To Beam Or Not To Beam: That is a Question of Cooperation for Language GANs". 2021

# SelfGAN

- Training via **imitation learning** using cooperatively generated texts
- Good results but unstable ⇒ require precise scheduler



SelfGAN



Discrete GAN

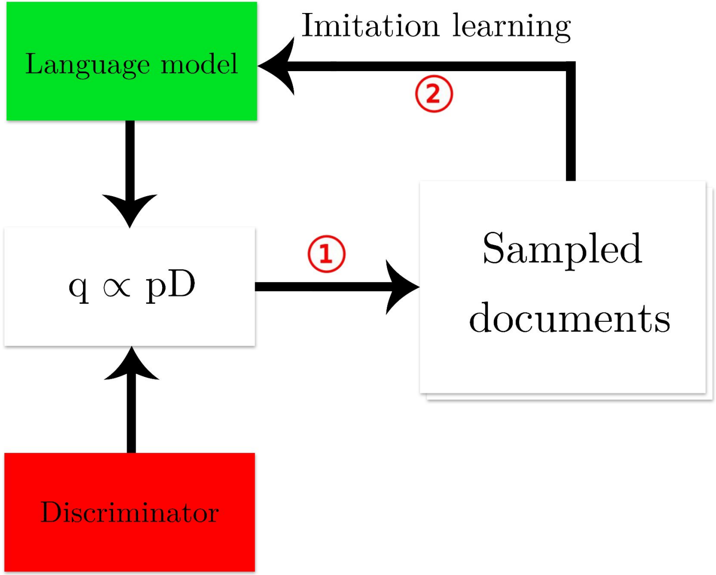
$$\nabla_\theta L_\theta \underset{\text{Cooperative sample}}{(x)} = -\nabla_\theta \overset{\text{Likelihood}}{\log p_\theta(x)}$$

$$\nabla_\theta L_\theta \underset{\text{Sample from the generator}}{(x)} = -\overset{\text{Discriminator score}}{D(x)}\,\nabla_\theta \overset{\text{Likelihood}}{\log p_\theta(x)}$$

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano. "To Beam Or Not To Beam: That is a Question of Cooperation for Language GANs". 2021

# RML-GAN

- Sample sequences from a target distribution q
- Learn to reproduce these sequences



① Generate M samples from $q \propto pD$

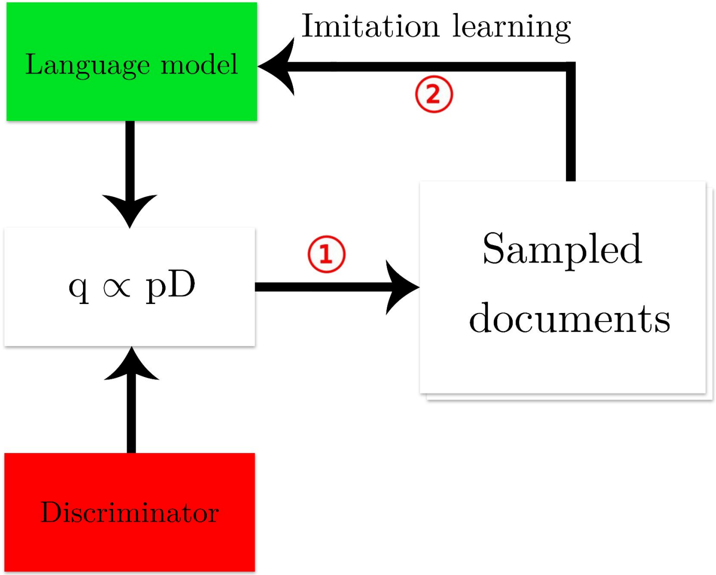
$$y^i \sim q(y^i)$$

② Train *p* using samples from *q*

$$\theta \leftarrow \theta + \epsilon \frac{1}{M} \sum_{i=1}^{M} \nabla_\theta \log p_\theta(y^i)$$

# RML-GAN

- Sample sequences from a target distribution q
- Learn to reproduce these sequences
- **Convergence guarantee if q = p*D**
  - … but we can’t sample from p*D



① Generate M samples from $q \propto pD$

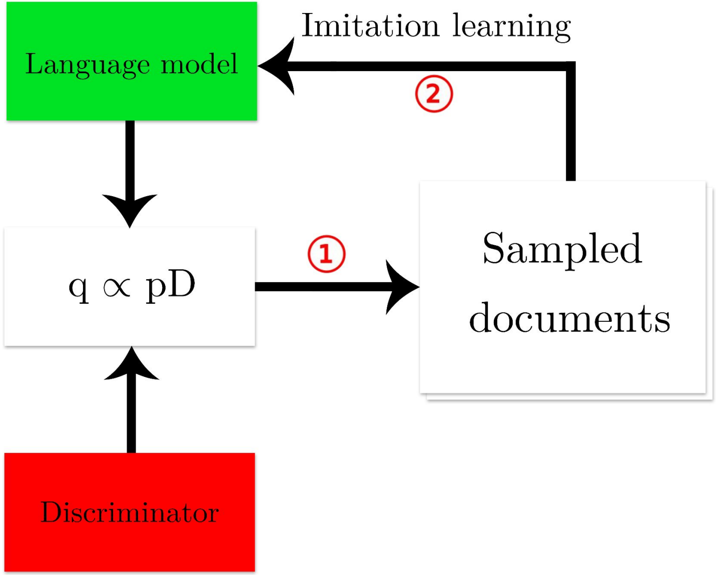
$$y^i \sim q(y^i)$$

② Train *p* using samples from *q*

$$\theta \leftarrow \theta + \epsilon \frac{1}{M} \sum_{i=1}^{M} \nabla_\theta \log p_\theta(y^i)$$

# RML-GAN

- Sample sequences from a target distribution q
- Learn to reproduce these sequences
- **Convergence guarantee if q = p*D**
  - … but we can’t sample from p*D
- ⇒ use importance sampling



① Generate M samples from $q \propto pD$

$$y^i \sim q(y^i)$$

② Train *p* using samples from *q*

$$\theta \leftarrow \theta + \epsilon \frac{1}{M} \sum_{i=1}^{M} \nabla_\theta \log p_\theta(y^i)$$

## Importance sampling

- We want to maximize

$$\mathbb{E}_{x\sim q(x)\propto p(x)D(x))}\left[\log p_\theta(x)\right]$$

# Importance sampling

- We want to maximize

$$\mathbb{E}_{x\sim q(x)\propto p(x)D(x))}\left[\log p_\theta(x)\right]$$

- Approximate expectation using an average

$$\mathbb{E}_{x\sim q(x)}\left[f(x)\right]\approx\frac{1}{N}\sum_{i=1}^{N}f(x_i)\qquad x_i\sim q(x)\ \text{(Monte Carlo Methods)}$$

# Importance sampling

- We want to maximize

$$\mathbb{E}_{x\sim q(x)\propto p(x)D(x))}\left[\log p_\theta(x)\right]$$

- Approximate expectation using an average

$$\mathbb{E}_{x\sim q(x)}\left[f(x)\right] \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i) \qquad x_i \sim q(x) \text{ (Monte Carlo Methods)}$$

- Introduce another (arbitrary) distribution

$$\mathbb{E}_{x\sim q(x)}\left[f(x)\right] = \int q(x)\left[f(x)\right]\,dx = \int \boxed{\frac{\hat{q}(x)}{\hat{q}(x)}} q(x)f(x)\,dx = \int \hat{q}(x)\left[\frac{q(x)}{\hat{q}(x)}f(x)\right]\,dx = \mathbb{E}_{x\sim\hat{q}(x)}\left[\frac{q(x)}{\hat{q}(x)}f(x)\right]$$

# Importance sampling

- We want to maximize

$$\mathbb{E}_{x \sim q(x) \propto p(x) D(x))} \left[ \log p_\theta(x) \right]$$

- Approximate expectation using an average

$$\mathbb{E}_{x \sim q(x)} \left[ f(x) \right] \approx \frac{1}{N} \sum_{i=1}^{N} f(x_i) \qquad x_i \sim q(x) \text{ (Monte Carlo Methods)}$$

- Introduce another (arbitrary) distribution

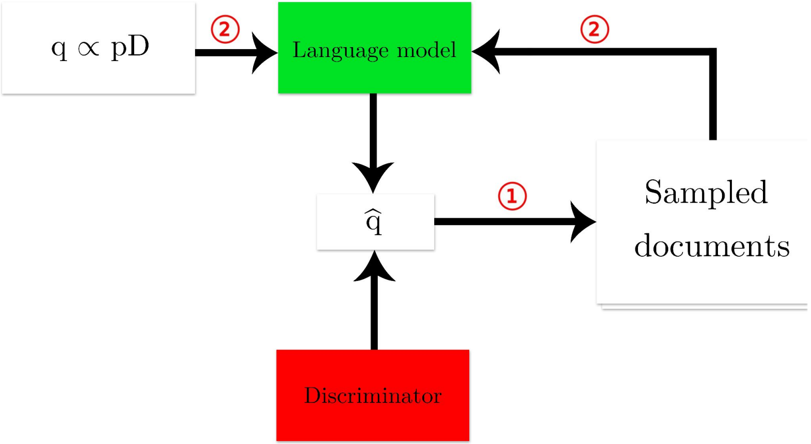
$$\mathbb{E}_{x \sim q(x)} \left[ f(x) \right] = \int q(x) \left[ f(x) \right] dx = \int \boxed{\frac{\hat{q}(x)}{\hat{q}(x)}} q(x) f(x) \, dx = \int \hat{q}(x) \left[ \frac{q(x)}{\hat{q}(x)} f(x) \right] dx = \mathbb{E}_{x \sim \hat{q}(x)} \left[ \frac{q(x)}{\hat{q}(x)} f(x) \right]$$

- Approximate expectation using samples from another distribution

$$\mathbb{E}_{x \sim q(x)} \left[ f(x) \right] \approx \frac{1}{N} \sum_{i=1}^{N} \boxed{\frac{q(x_i)}{\hat{q}(x_i)}} f(x_i) \qquad x_i \sim \hat{q}(x)$$

# Generative Cooperative Networks

- Sample from an **arbitrary distribution**
- Unbias the estimation of the gradient using **importance sampling weights**



① Generate M samples from $\hat{q}$

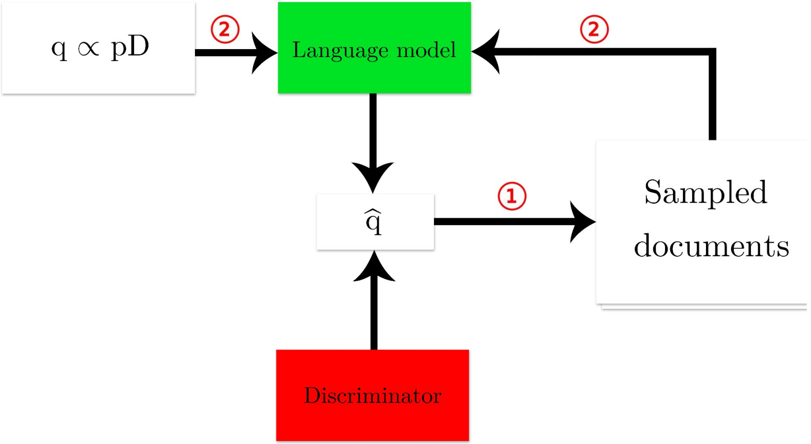
$$y^i \sim \hat{q}(y^i)$$

② Train p using weighted importance sampling

$$\theta \leftarrow \theta + \epsilon \frac{1}{\sum_{i=1}^{M} w^i} \sum_{i=1}^{M} w^i \nabla_\theta \log p_\theta(y^i) \quad \text{with: } w^i = \frac{q(y^i)}{\hat{q}(y^i)}$$

# Generative Cooperative Networks (GCN)

- Sample from an **arbitrary distribution**
- Unbias the estimation of the gradient using **importance sampling weights**
- Traditional GAN update up to a **normalization factor**
    - **Automatic scheduler**



① Generate M samples from $\hat{q}$

$$y^i \sim \hat{q}(y^i)$$

② Train p using weighted importance sampling

$$\theta \leftarrow \theta + \epsilon \frac{1}{\sum_{i=1}^{M} w^i} \sum_{i=1}^{M} w^i \nabla_\theta \log p_\theta(y^i) \quad \text{with: } w^i = \frac{q(y^i)}{\hat{q}(y^i)}$$

$$\text{With } \hat{q} = p_\theta \text{, we have: } \theta \leftarrow \theta + \epsilon \underbrace{\frac{1}{Z_t}}_{\text{Normalization term}} \underbrace{\sum_{i=1}^{M} D_t\left(y^i\right) \nabla_\theta \log p_\theta\left(y^i\right)}_{\text{Text GAN update}} \text{, where } Z_t = \sum_{i=1}^{M} D_t\left(y^i\right)$$

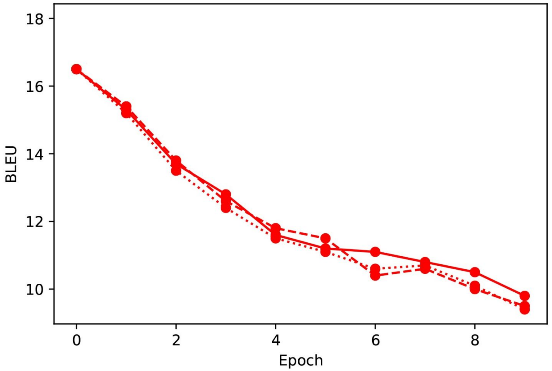
# Results

- Models **diverge without scheduler**

⋯ $\hat{q} = p$

– – – $\hat{q} = Nucleus$

—— $\hat{q} = MCTS$

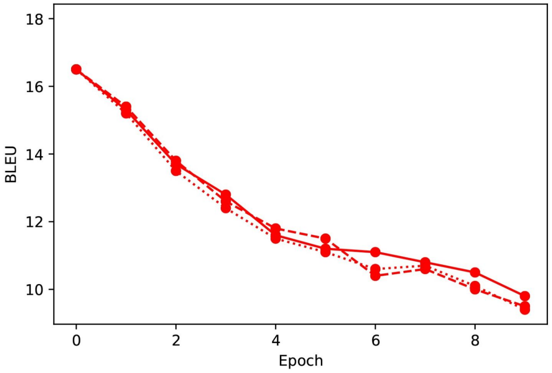


No scheduler

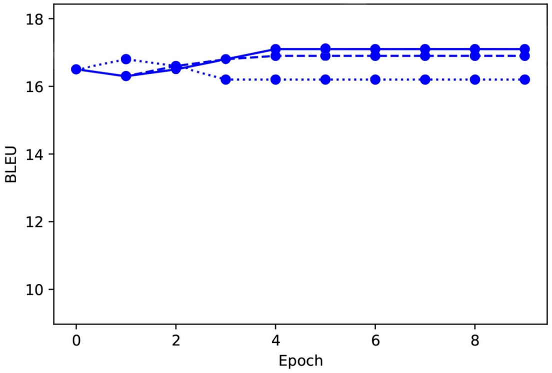
# Results

- Models **diverge without scheduler**
- Scheduler prevents divergence

- ··· $\hat{q} = p$
- ▬ ▬ ▬ $\hat{q} = Nucleus$
- ▬▬▬ $\hat{q} = MCTS$



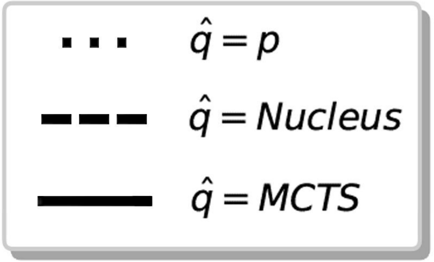
No scheduler

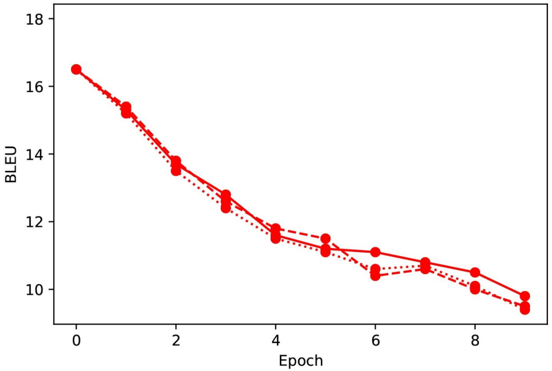


Manual scheduler

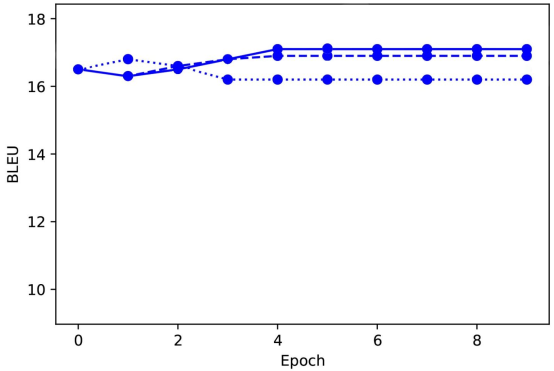
# Results

- Models **diverge without scheduler**
- Scheduler prevents divergence
- Normalization term replaces the scheduler

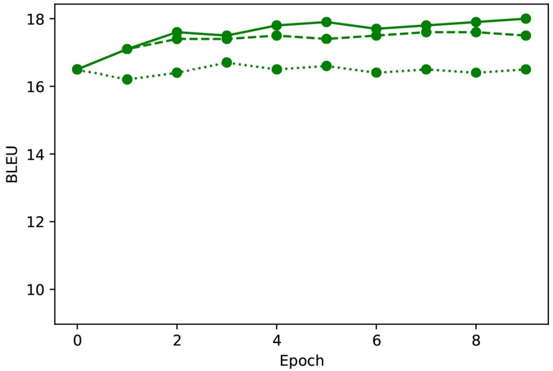




No scheduler



Manual scheduler



Normalization (Z)

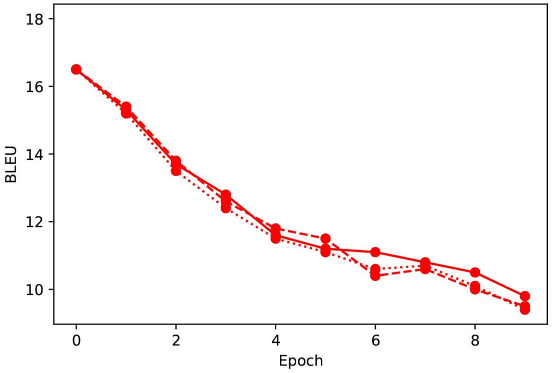
# Results

- Models **diverge without scheduler**
- Scheduler prevents divergence
- Normalization term replaces the scheduler
- Sampling closer to q further improves the results

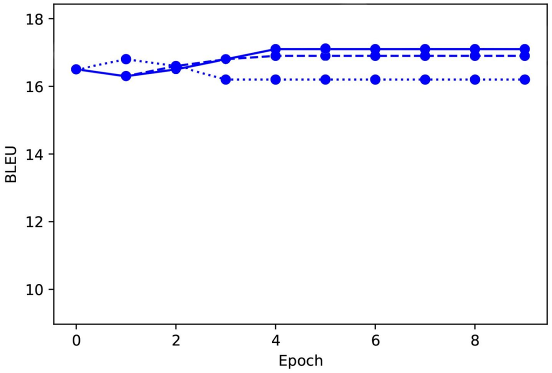
··· $\hat{q} = p$

– – – $\hat{q} = Nucleus$

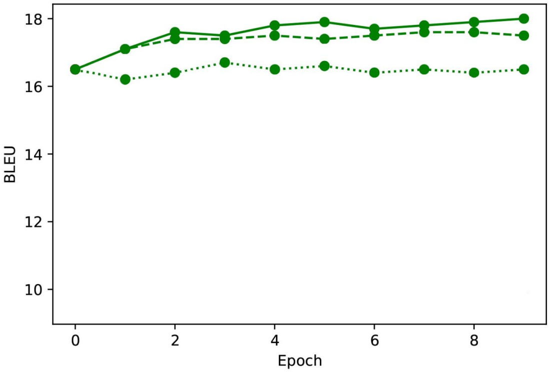
—— $\hat{q} = MCTS$



No scheduler



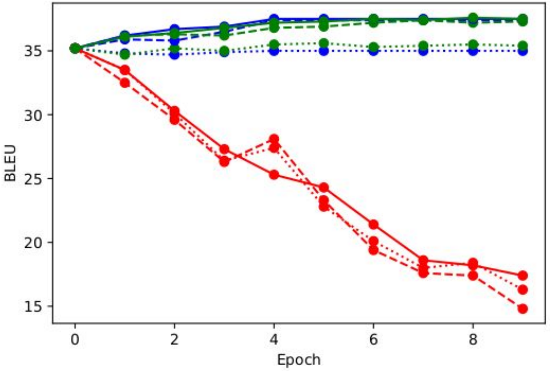
Manual scheduler



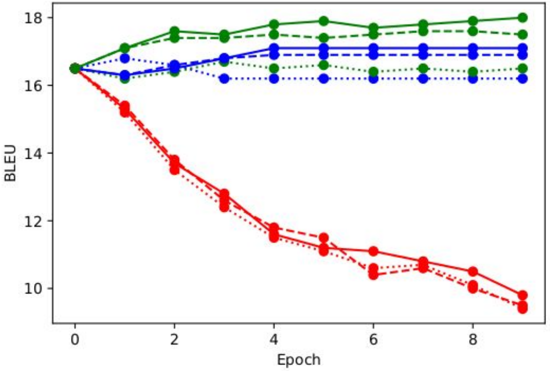
Normalization (Z)

# Results

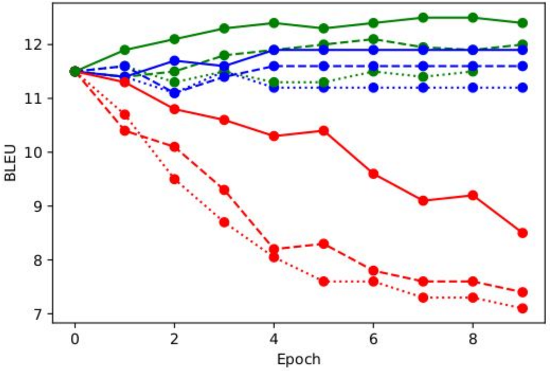
- **GCN+MCTS yields state-of-art results on 2 tasks without any scheduler**



Unconditional NLG



Question Generation



Summarization

···  $\hat{q} = p$

- - -  $\hat{q} = Nucleus$

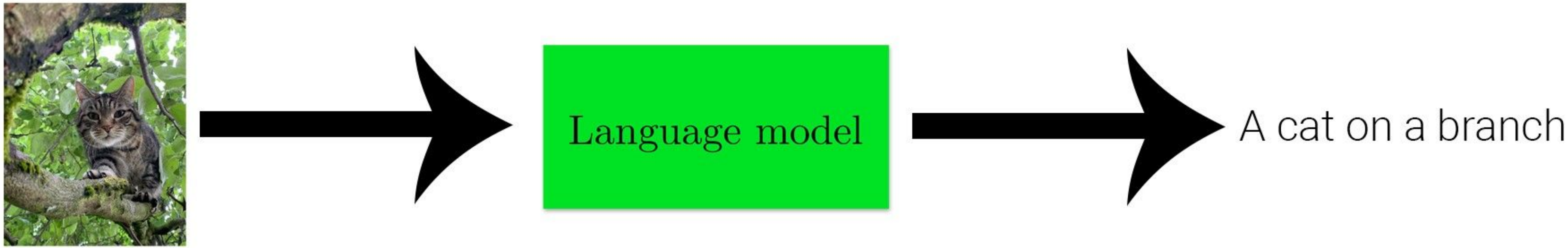
——  $\hat{q} = MCTS$

- No scheduler
- Manual scheduler (SOTA)
- Normalization (Z)

1. PPL-MCTS: Constrained Textual Generation Through Discriminator-Guided MCTS Decoding
2. Generative Cooperative Networks for Natural Language Generation
3. **Distinctive Image Captioning: Leveraging Ground Truth Captions in CLIP Guided Reinforcement Learning**
4. Conclusion & perspectives

# Image Captioning

- Language model conditioned on an image
- Create a **powerful cross-modal alignment**[1]



Language model

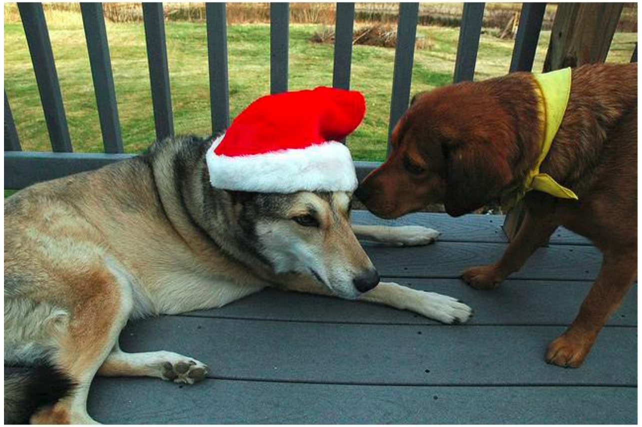
A cat on a branch

[1] Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, Lucas Beyer. “Image Captioners Are Scalable Vision Learners Too”. 2023

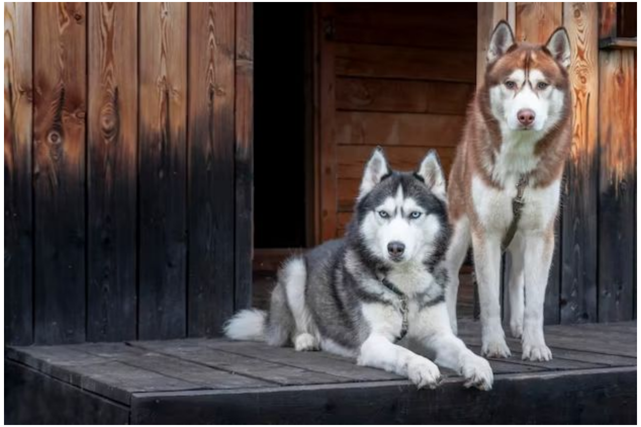
# Distinctive Image Captioning

- Datasets captions only describe most salient objects, common to many images
- Higher word-matching metrics with words common across different images, not specific ones

*A couple of dogs standing on a porch*



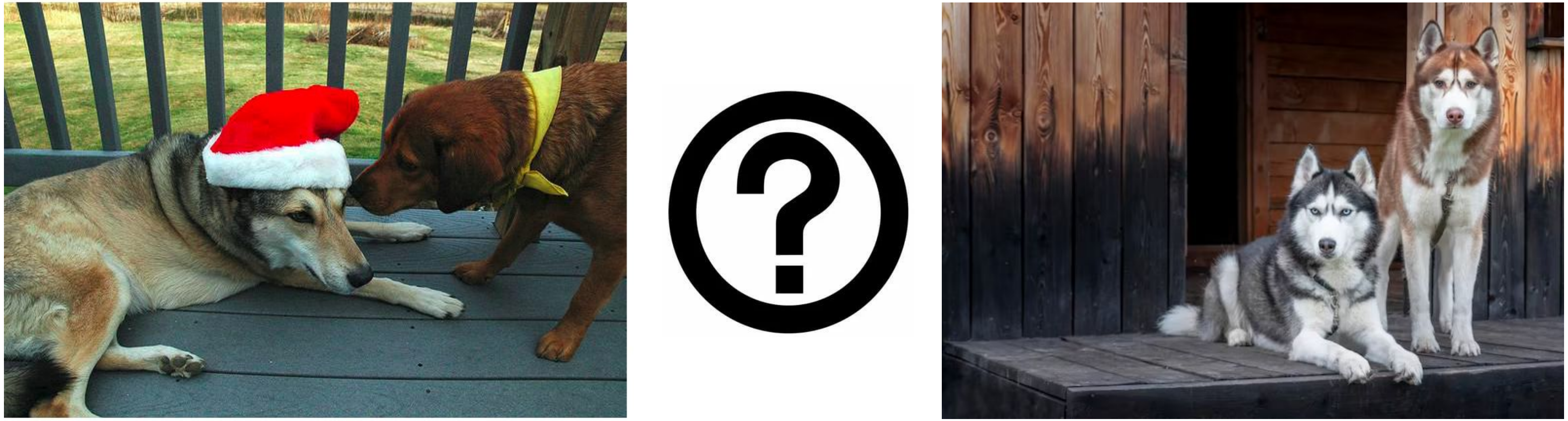
?



# Distinctive Image Captioning

- Datasets captions only describe most salient objects, common to many images
- Higher word-matching metrics with words common across different images, not specific ones

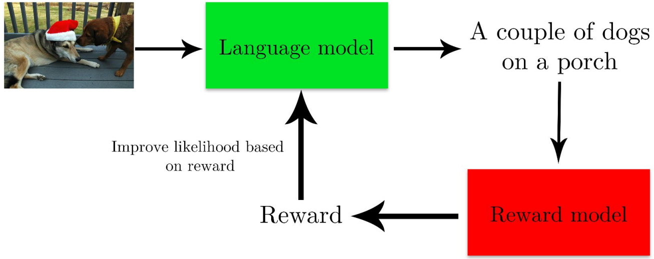
*A couple of dogs standing on a porch*



- Fine-grained alignment to describe **this image and only this one**

# Cross-modal retriever rewards

- Reinforcement learning to optimize cross-modal similarity of the generated caption and the target image
  - A description that can let the retriever identify the image

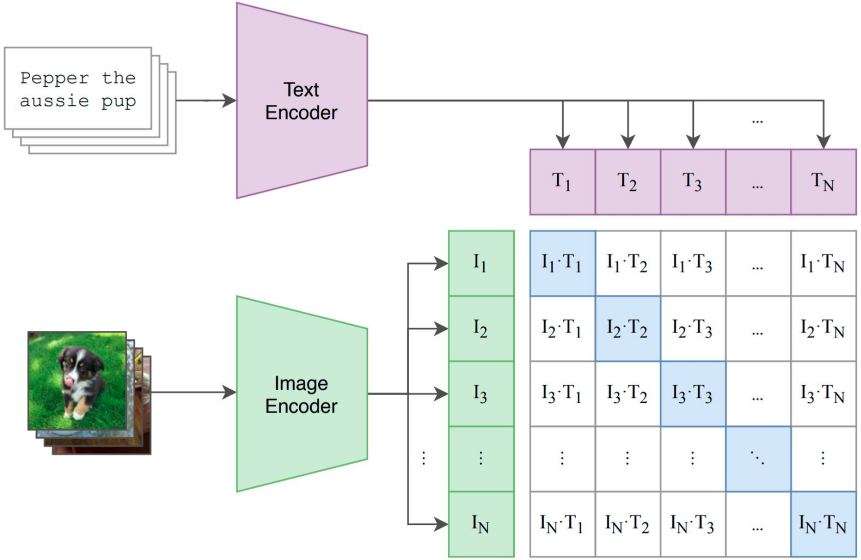

Language model
A couple of dogs on a porch
Reward model
Reward
Improve likelihood based on reward

$$\nabla_\theta L_\theta \underbrace{(x)}_{\text{Sample from the generator}} = -\overbrace{r_{sim}(x)}^{\text{Similarity reward}} \nabla_\theta \underbrace{\log p_\theta (x)}_{\text{Likelihood}}$$

*a couple of dogs wearing a santa hat on a porch*

# Contrastive learning

- Dual encoder, each projecting a modality separately
  - Similarity using dot product of both representations

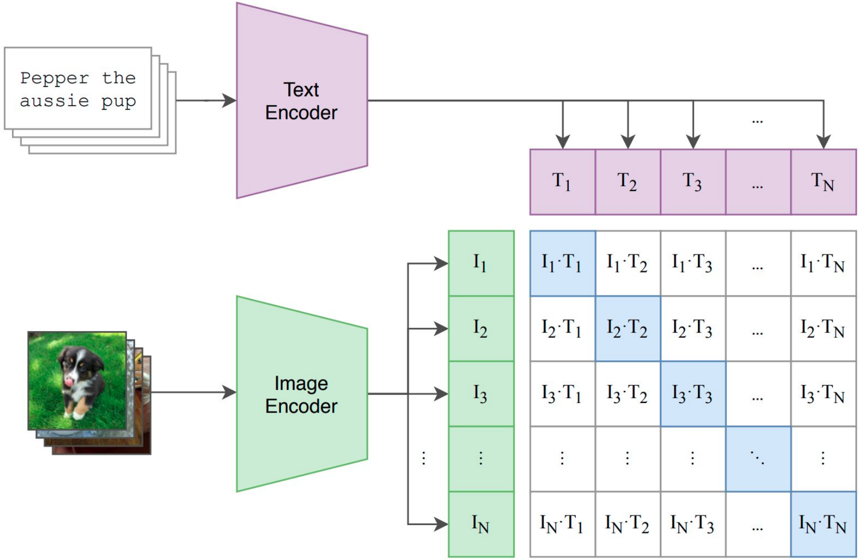


|  | $T_1$ | $T_2$ | $T_3$ | ... | $T_N$ |
|---|---|---|---|---|---|
| $I_1$ | $I_1 \cdot T_1$ | $I_1 \cdot T_2$ | $I_1 \cdot T_3$ | ... | $I_1 \cdot T_N$ |
| $I_2$ | $I_2 \cdot T_1$ | $I_2 \cdot T_2$ | $I_2 \cdot T_3$ | ... | $I_2 \cdot T_N$ |
| $I_3$ | $I_3 \cdot T_1$ | $I_3 \cdot T_2$ | $I_3 \cdot T_3$ | ... | $I_3 \cdot T_N$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $I_N$ | $I_N \cdot T_1$ | $I_N \cdot T_2$ | $I_N \cdot T_3$ | ... | $I_N \cdot T_N$ |

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever. “Learning Transferable Visual Models From Natural Language Supervision”. 2021

# Contrastive learning

- Dual encoder, each projecting a modality separately
  - Similarity using dot product of both representations
- Couple closer than any element in the batch



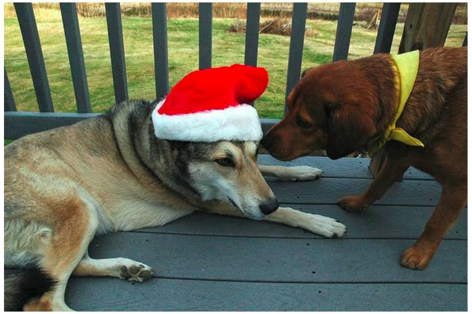
|  | $T_1$ | $T_2$ | $T_3$ | ... | $T_N$ |
|---|---|---|---|---|---|
| $I_1$ | $I_1{\cdot}T_1$ | $I_1{\cdot}T_2$ | $I_1{\cdot}T_3$ | ... | $I_1{\cdot}T_N$ |
| $I_2$ | $I_2{\cdot}T_1$ | $I_2{\cdot}T_2$ | $I_2{\cdot}T_3$ | ... | $I_2{\cdot}T_N$ |
| $I_3$ | $I_3{\cdot}T_1$ | $I_3{\cdot}T_2$ | $I_3{\cdot}T_3$ | ... | $I_3{\cdot}T_N$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ |
| $I_N$ | $I_N{\cdot}T_1$ | $I_N{\cdot}T_2$ | $I_N{\cdot}T_3$ | ... | $I_N{\cdot}T_N$ |

$$\mathcal{L}_{\text{CLIP}} = \underbrace{\log \frac{e^{\frac{t_c \cdot i_c}{\tau}}}{\sum_{t\in\mathcal{T}} e^{\frac{t \cdot i_c}{\tau}}}}_{\text{image-to-text}} + \underbrace{\log \frac{e^{\frac{t_c \cdot i_c}{\tau}}}{\sum_{i\in\mathcal{I}} e^{\frac{t_c \cdot i}{\tau}}}}_{\text{text-to-image}}$$

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever. “Learning Transferable Visual Models From Natural Language Supervision”. 2021

# Regularization

- Prevent the model from learning ill-formed solutions



*a close up of two* ***brown*** *and* ***black dogs*** *wearing a* ***santa hat*** *on a* ***black*** *and* ***brown dog*** *with a* ***red hat*** *on a backyard with a fence in the background*

# Regularization

- Prevent the model from learning ill-formed solutions
- Regularization term in the reward
  - ○ KL divergence, CIDEr value, **grammar network**…

*a close up of two **brown** and **black dogs** wearing a **santa hat** on a **black** and **brown dog** with a **red hat** on a backyard with a fence in the background*

$$\nabla_\theta L_\theta \underbrace{(x)}_{\text{Sample from the generator}} = -\left[\left(\overbrace{\alpha\, r_{sim}(x)}^{\text{Similarity reward}} + \overbrace{(1-\alpha)\, r_{regu}(x)}^{\text{Regularization reward}}\right)\nabla_\theta \underbrace{\log p_\theta\,(x)}_{\text{Likelihood}}\right]$$

# Experiments

- 3 different contributions to improve CLIP-based RL image captioning
    1. **Discriminator regularization**
    2. **RL objective on ground truth samples**
    3. **Bidirectional contrastive reward**

# Experiments

- 3 different contributions to improve CLIP-based RL image captioning
  1. **Discriminator regularization**
  2. **RL objective on ground truth samples**
  3. **Bidirectional contrastive reward**
- MS COCO dataset
- Trade-off:
  - **Discriminativeness**: recall@k using generated caption (fixed CLIP model)
  - **Writing quality**: BLEU, ROUGE, CIDEr, METEOR and SPICE

# Discriminator regularization

- Use generated text discriminator scores as regularization
- Simple MLP using CLIP representations as input

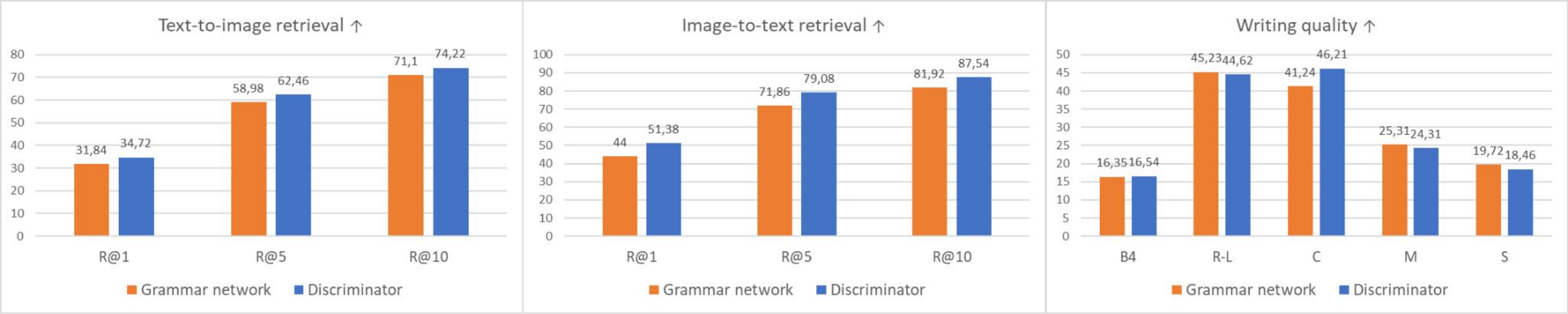
$$\nabla_\theta L_\theta \underbrace{(x)}_{\text{Sample from the generator}} = -\left[\left(\overbrace{\alpha\, r_{sim}(x)}^{\text{Similarity reward}} + \overbrace{(1-\alpha)\, r_{regu}(x)}^{\text{Regularization reward}}\right)\nabla_\theta \underbrace{\log p_\theta\,(x)}_{\text{Likelihood}}\right]$$

# Discriminator regularization

- Use generated text discriminator scores as regularization
- Simple MLP using CLIP representations as input

$$\nabla_\theta L_\theta \underbrace{(x)}_{\text{Sample from the generator}} = -\left[\left(\overbrace{\alpha\, r_{sim}(x)}^{\text{Similarity reward}} + \overbrace{(1-\alpha)\, r_{regu}(x)}^{\text{Regularization reward}}\right) \nabla_\theta \underbrace{\log p_\theta\,(x)}_{\text{Likelihood}}\right]$$

- Higher retrieval rate without degrading written quality



# Weighted Teacher Forcing

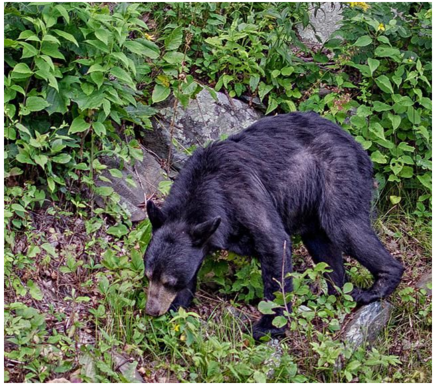
- RL learns from high-scoring sequences
- Ground truths are (relatively) good solutions

$$\nabla_\theta L_\theta \underbrace{(x)}_{\text{Ground truth sample}} = -\left( \overbrace{\alpha \, r_{sim}(x)}^{\text{Similarity reward}} + \overbrace{(1-\alpha)\, r_{regu}(x)}^{\text{Regularization reward}} \right) \nabla_\theta \underbrace{\log p_\theta \, (x)}_{\text{Likelihood}}$$

# Weighted Teacher Forcing

- RL learns from high-scoring sequences
- Ground truths are (relatively) good solutions
- Learn to reproduce human-written sequence (TF) but focuses on highly descriptive ones

$$\nabla_\theta L_\theta \underbrace{(x)}_{\text{Ground truth sample}} = -\left( \overbrace{\alpha\, r_{sim}(x)}^{\text{Similarity reward}} + \overbrace{(1-\alpha)\, r_{regu}(x)}^{\text{Regularization reward}} \right) \nabla_\theta \underbrace{\log p_\theta (x)}_{\text{Likelihood}}$$

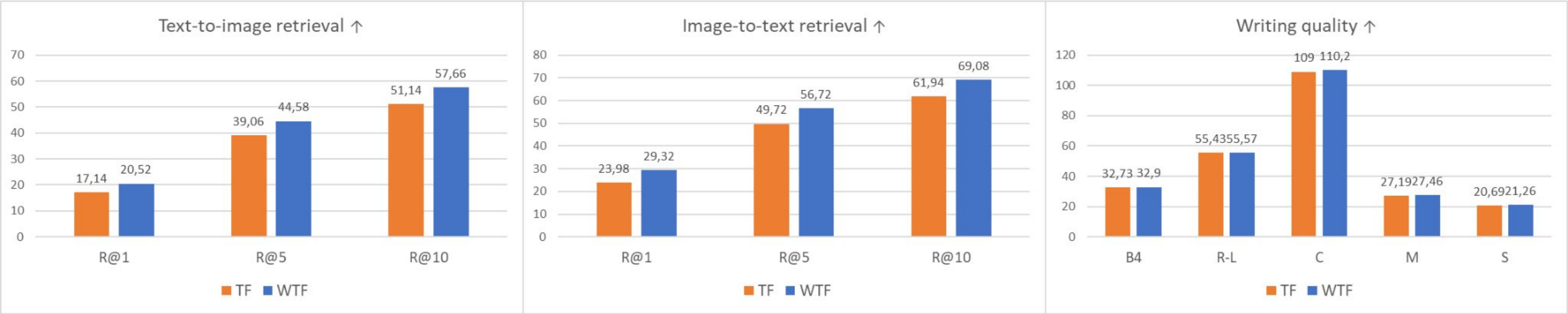


*(1) there is an adult bear that is walking in the forest*

*(2) picture of an exterior place that looks wonderful.*

# Weighted Teacher Forcing

- Improve retrieval metrics using only ground truth, without degrading writing quality
- Better regularization objective to couple with traditional RL



# Baseline rewards

- Subtract a baseline to the reward to reduce variance

$$\nabla_\theta L_\theta \underset{\substack{\uparrow \\ \text{Sample from the generator}}}{(x)} = -\left(\overset{\substack{\text{Reward} \\ \downarrow}}{r(x)} - \overset{\substack{\text{Baseline} \\ \downarrow}}{b}\right)\nabla_\theta \underset{\substack{\uparrow \\ \text{Likelihood}}}{\log p_\theta\,(x)}$$

# Baseline rewards

- Subtract a baseline to the reward to reduce variance

$$\nabla_\theta L_\theta \underset{\text{Sample from the generator}}{(x)} = -\left( \overset{\text{Reward}}{r(x)} - \overset{\text{Baseline}}{b} \right) \nabla_\theta \underset{\text{Likelihood}}{\log p_\theta (x)}$$

1. Use the model itself as a baseline[1]

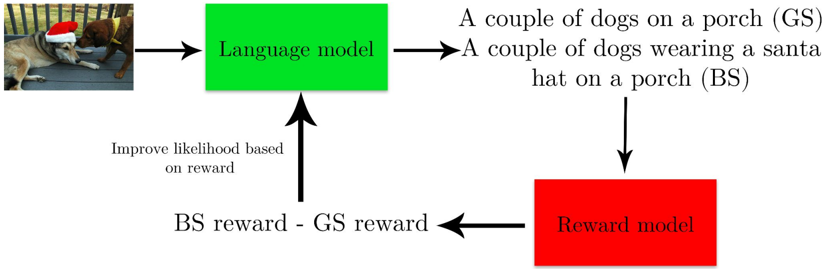

Language model
A couple of dogs on a porch (GS)
A couple of dogs wearing a santa hat on a porch (BS)
Reward model
BS reward - GS reward
Improve likelihood based on reward

Image-to-text baseline

[1] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, Mohit Bansal. “Fine-grained Image Captioning with CLIP Reward”. 2022

# Baseline rewards

- Subtract a baseline to the reward to reduce variance

$$\nabla_\theta L_\theta \underbrace{(x)}_{\text{Sample from the generator}} = -\left(\overbrace{r(x)}^{\text{Reward}} - \overbrace{b}^{\text{Baseline}}\right)\nabla_\theta \underbrace{\log p_\theta(x)}_{\text{Likelihood}}$$

1. Use the model itself as a baseline[1]
2. Similarity with other (similar) images[2]

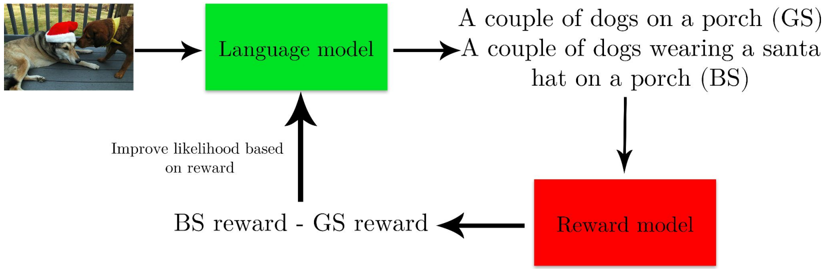
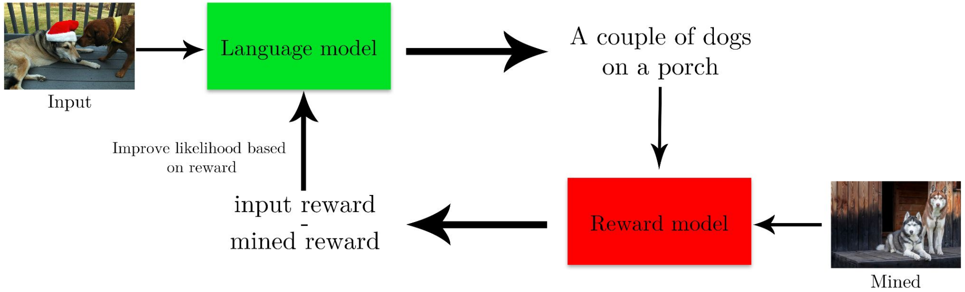

Language model
A couple of dogs on a porch (GS)
A couple of dogs wearing a santa hat on a porch (BS)
Reward model
BS reward - GS reward
Improve likelihood based on reward

Image-to-text baseline


Input
Language model
A couple of dogs on a porch
Reward model
Mined
input reward
-
mined reward
Improve likelihood based on reward

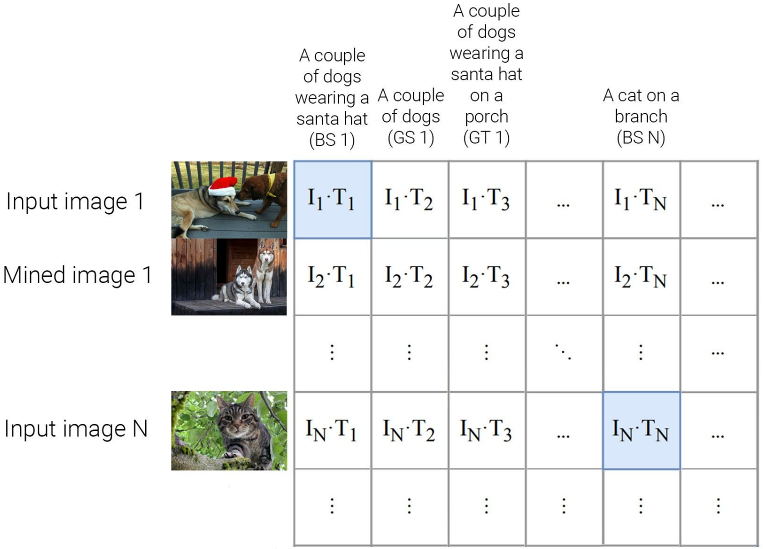
Text-to-image baseline

[1] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, Mohit Bansal. "Fine-grained Image Captioning with CLIP Reward". 2022

[2] Youyuan Zhang, Jiuniu Wang, Hao Wu, Wenjia Xu. "Distinctive Image Captioning via CLIP Guided Group Optimization". 2022

# Bidirectional Contrastive Reward

- Decoupled contrastive loss



| | A couple of dogs wearing a santa hat (BS 1) | A couple of dogs (GS 1) | A couple of dogs wearing a santa hat on a porch (GT 1) | | A cat on a branch (BS N) | |
|---|---|---|---|---|---|---|
| Input image 1 | $I_1 \cdot T_1$ | $I_1 \cdot T_2$ | $I_1 \cdot T_3$ | ... | $I_1 \cdot T_N$ | ... |
| Mined image 1 | $I_2 \cdot T_1$ | $I_2 \cdot T_2$ | $I_2 \cdot T_3$ | ... | $I_2 \cdot T_N$ | ... |
| | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ... |
| Input image N | $I_N \cdot T_1$ | $I_N \cdot T_2$ | $I_N \cdot T_3$ | ... | $I_N \cdot T_N$ | ... |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

$$r_{bicont}(t_c) = \tau \left( \underbrace{\log \frac{e^{\frac{t_c \cdot i_c}{\tau}}}{\sum_{t \in \mathcal{T} \setminus t_c} e^{\frac{t \cdot i_c}{\tau}}}}_{\text{Image-to-text reward } r_{i2t}(t_c)} + \underbrace{\log \frac{e^{\frac{t_c \cdot i_c}{\tau}}}{\sum_{i \in \mathcal{I} \setminus i_c} e^{\frac{t_c \cdot i}{\tau}}}}_{\text{Text-to-image reward } r_{t2i}(t_c)} \right)$$

# Bidirectional Contrastive Reward

- Decoupled contrastive loss
- Closest element in the batch as baseline
- Natively handle both cross-modal directions

| | A couple of dogs wearing a santa hat (BS 1) | A couple of dogs (GS 1) | A couple of dogs wearing a santa hat on a porch (GT 1) | | A cat on a branch (BS N) | |
|---|---|---|---|---|---|---|
| Input image 1 | $I_1 \cdot T_1$ | $I_1 \cdot T_2$ | $I_1 \cdot T_3$ | ... | $I_1 \cdot T_N$ | ... |
| Mined image 1 | $I_2 \cdot T_1$ | $I_2 \cdot T_2$ | $I_2 \cdot T_3$ | ... | $I_2 \cdot T_N$ | ... |
| | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ... |
| Input image N | $I_N \cdot T_1$ | $I_N \cdot T_2$ | $I_N \cdot T_3$ | ... | $I_N \cdot T_N$ | ... |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

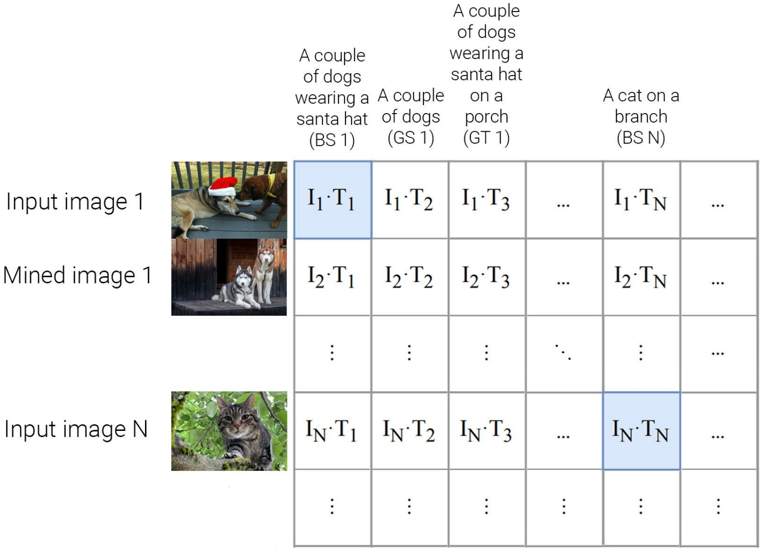
$$r_{bicont}(t_c) = \tau \left( \underbrace{\log \frac{e^{\frac{t_c \cdot i_c}{\tau}}}{\sum_{t \in \mathcal{T} \setminus t_c} e^{\frac{t \cdot i_c}{\tau}}}}_{\text{Image-to-text reward } r_{i2t}(t_c)} + \underbrace{\log \frac{e^{\frac{t_c \cdot i_c}{\tau}}}{\sum_{i \in \mathcal{I} \setminus i_c} e^{\frac{t_c \cdot i}{\tau}}}}_{\text{Text-to-image reward } r_{t2i}(t_c)} \right)$$

$$r_{i2t}(t_c) = \tau \left( \log \left( e^{\frac{t_c \cdot i_c}{\tau}} \right) - \log \left( \sum_{t \in \mathcal{T} \setminus t_c} e^{\frac{t \cdot i_c}{\tau}} \right) \right)$$

$$\approx t_c \cdot i_c - \max_{t \in \mathcal{T} \setminus t_c} \{ t \cdot i_c \}$$

# Bidirectional Contrastive Reward

- Decoupled contrastive loss
- Closest element in the batch as baseline
- Natively handle both cross-modal directions
- The caption is **very descriptive of the image and this image only**



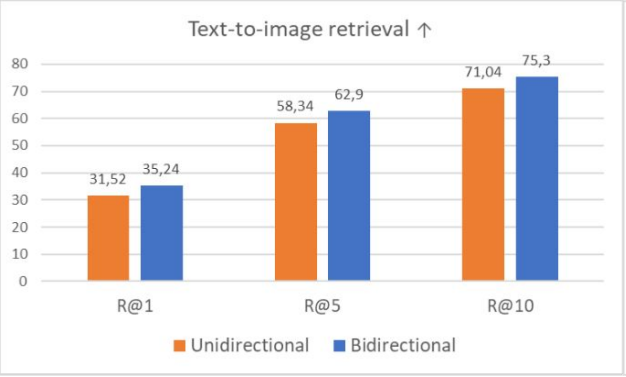
| | A couple of dogs wearing a santa hat (BS 1) | A couple of dogs (GS 1) | A couple of dogs wearing a santa hat on a porch (GT 1) | | A cat on a branch (BS N) | |
|---|---|---|---|---|---|---|
| Input image 1 | $I_1 \cdot T_1$ | $I_1 \cdot T_2$ | $I_1 \cdot T_3$ | ... | $I_1 \cdot T_N$ | ... |
| Mined image 1 | $I_2 \cdot T_1$ | $I_2 \cdot T_2$ | $I_2 \cdot T_3$ | ... | $I_2 \cdot T_N$ | ... |
| | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ... |
| Input image N | $I_N \cdot T_1$ | $I_N \cdot T_2$ | $I_N \cdot T_3$ | ... | $I_N \cdot T_N$ | ... |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

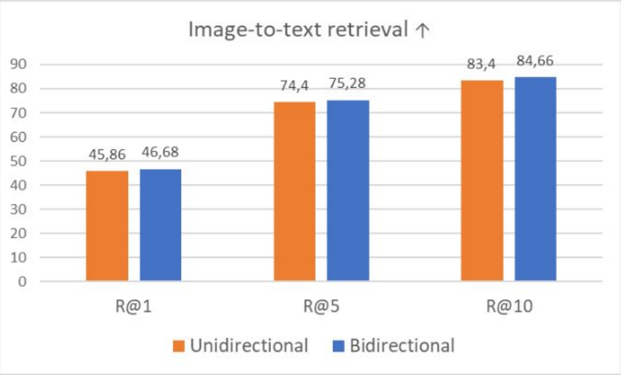
$$r_{bicont}(t_c) = \tau \left( \underbrace{\log \frac{e^{\frac{t_c \cdot i_c}{\tau}}}{\sum_{t \in \mathcal{T} \setminus t_c} e^{\frac{t \cdot i_c}{\tau}}}}_{\text{Image-to-text reward } r_{i2t}(t_c)} + \underbrace{\log \frac{e^{\frac{t_c \cdot i_c}{\tau}}}{\sum_{i \in \mathcal{I} \setminus i_c} e^{\frac{t_c \cdot i}{\tau}}}}_{\text{Text-to-image reward } r_{t2i}(t_c)} \right)$$

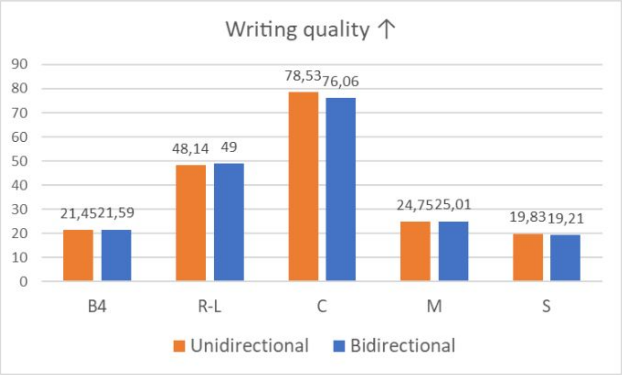
$$\begin{aligned} r_{i2t}(t_c) &= \tau \left( \log \left( e^{\frac{t_c \cdot i_c}{\tau}} \right) - \log \left( \sum_{t \in \mathcal{T} \setminus t_c} e^{\frac{t \cdot i_c}{\tau}} \right) \right) \\ &\approx t_c \cdot i_c - \max_{t \in \mathcal{T} \setminus t_c} \{ t \cdot i_c \} \end{aligned}$$

# Bidirectional reward

- Unidirectional image-to-text reward only yield significantly lower text-to-image retrieval results
- Both cross-modal directions are needed for a caption highly descriptive of **this image and this image only**







1. PPL-MCTS: Constrained Textual Generation Through Discriminator-Guided MCTS Decoding
2. Generative Cooperative Networks for Natural Language Generation
3. Distinctive Image Captioning: Leveraging Ground Truth Captions in CLIP Guided Reinforcement Learning
4. **Conclusion & perspectives**

# Take-home message

1. Cooperative generation using **MCTS to search** in the LM space[1]
   - ○ State-of-the-art on **constraint generation without tuning the LM**
2. Novel formulation of textual GANs **with convergence guarantee**[2]
   - ○ State-of-the-art on **natural language generation** using cooperative generation
3. CLIP-guided cross-modal generation to create **distinctive captions**[3]
   - ○ Use GT to **ground learning into human distribution**
   - ○ Consider **both retrieval directions** to create truly distinctive captions

[1] **Antoine Chaffin**, Vincent Claveau, Ewa Kijak. "PPL-MCTS: Constrained Textual Generation Through Discriminator-Guided MCTS Decoding". *NAACL-HLT*. 2022
[2] Sylvain Lamprier, Thomas Scialom, **Antoine Chaffin**, Vincent Claveau, Ewa Kijak, Jacopo Staiano, Benjamin Piwowarski. "Generative Cooperative Networks for Natural Language Generation". *ICML*. 2022
[3] **Antoine Chaffin**, Vincent Claveau, Ewa Kijak. "Distinctive Image Captioning: Leveraging Ground Truth Captions in CLIP Guided Reinforcement Learning". *Under review*. 2023

# Perspectives

- **Synthetic data yield strong performances**
  - **DALL-E 3, Unnatural Code Llama, BLIP-2**
- **Cooperative generation allows to tweak the generation with explicit constraint**
  - **Many use cases**: Textual GANs, RLHF, texts representative of the guide (explainability)★, factual consistency, logical constraint, unit tests, additional performance, numerical planning, steganography…
  - Computational overhead★
- **Multimodal alignment using cross-modal generation**
  - **Extend our NLP work to multimodal data**

★ **Antoine Chaffin**, Julien Delaunay. "``Honey, Tell Me What's Wrong'', Global Explainability of Textual Discriminative Models through Cooperative Generation". *BlackBoxNLP*. 2023

★ **Antoine Chaffin**, Thomas Scialom, Sylvain Lamprier, Jacopo Staiano, Benjamin Piwowarski, Ewa Kijak, Vincent Claveau. "Which Discriminator for Cooperative Text Generation?". *SIGIR*. 2022

# Generative Models and Misinformation: A Double-Edged Sword

- **Growing capacity and accessibility of generative models**
- **Lots of opportunities**
  - General-purpose language models
  - **Retrieval-augmented language model**
- **Lots of risks**
  - Humans are easily fooled
  - **Automatic detection is hard (impossible?)**
  - Other modalities
- **Watermarking can mitigate the issue**
  - Need to be integrated **into** the model
  - Emerging research domain★
  - Reverse the paradigm: only trust watermarked content

★ Pierre Fernandez, **Antoine Chaffin**, Karim Tit, Vivien Chappelier, Teddy Furon. "Three Bricks to Consolidate Watermarks for Large Language Models". *WIFS*. 2023

# Contributions

## International publications

- Vincent Claveau, **Antoine Chaffin**, Ewa Kijak. "Generating artificial texts as substitution or complement of training data". Proceedings of the 2022 Language Resources and Evaluation Conference (LREC). 2022
- **Antoine Chaffin**, Vincent Claveau, Ewa Kijak. "PPL-MCTS: Constrained Textual Generation Through Discriminator-Guided MCTS Decoding". Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). 2022
- Sylvain Lamprier, Thomas Scialom, **Antoine Chaffin**, Vincent Claveau, Ewa Kijak, Jacopo Staiano, Benjamin Piwowarski. "Generative Cooperative Networks for Natural Language Generation". Proceedings of the 39th International Conference on Machine Learning (ICML). 2022
- **Antoine Chaffin**, Thomas Scialom, Sylvain Lamprier, Jacopo Staiano, Benjamin Piwowarski, Ewa Kijak, Vincent Claveau. "Which Discriminator for Cooperative Text Generation?". Proceedings of the 45th International Conference on Research and Development in Information Retrieval (SIGIR). 2022
- Pierre Fernandez, **Antoine Chaffin**, Karim Tit, Vivien Chappelier, Teddy Furon. "Three Bricks to Consolidate Watermarks for Large Language Models." Proceeding of the 2023 International Workshop on Information Forensics and Security (WIFS). 2023
- **Antoine Chaffin**, Julien Delaunay. "``Honey, Tell Me What's Wrong'', Global Explainability of Textual Discriminative Models through Cooperative Generation". Proceedings of the Sixth Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP). 2023

## National publications

- Vincent Claveau, **Antoine Chaffin**, Ewa Kijak. "La génération de textes artificiels en substitution ou en complément de données d’apprentissage". Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles (TALN). 2021
- **Antoine Chaffin**, Vincent Claveau, Ewa Kijak. "Décodage guidé par un discriminateur avec le Monte Carlo Tree Search pour la génération de texte contrainte". Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles (TALN). 2022
- **Antoine Chaffin**, Thomas Scialom, Sylvain Lamprier, Jacopo Staiano, Benjamin Piwowarski, Ewa Kijak, Vincent Claveau. "Choisir le bon co-équipier pour la génération coopérative de texte". Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles (TALN). 2022
- **Antoine Chaffin**, Julien Delaunay. "``Honey, Tell Me What's Wrong'', Explicabilité Globale des Modèles de TAL par la Génération Coopérative. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN). 2023

## Under review

- **Antoine Chaffin**, Vincent Claveau, Ewa Kijak. "Distinctive Image Captioning: Leveraging Ground Truth Captions in CLIP Guided Reinforcement Learning". Under review. 2023

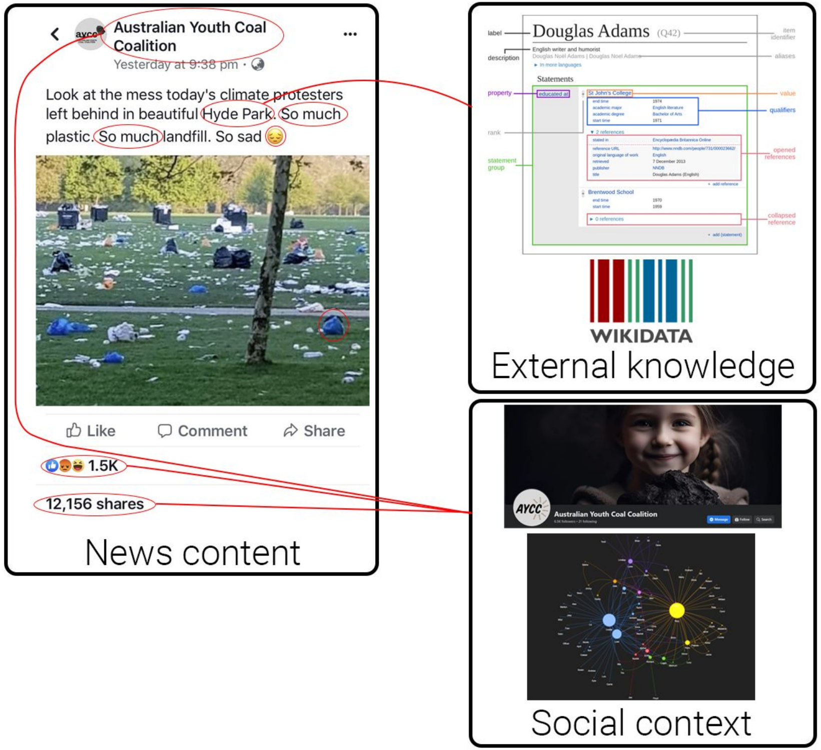
## Big Science

- Victor Sanh et al. “Multitask Prompted Training Enables Zero-Shot Task Generalization”. Proceedings of the 2022 International Conference on Learning Representations (ICLR). 2022
- BigScience Workshop. “BLOOM: A 176B-Parameter Open-Access Multilingual Language Model”. ArXiv. 2023

# Director’s cut

# Automatic misinformation detection

- Manual fact-checking too long and costly
- Automatic detection
  - Social context analysis
  - Statistical differences
  - Semantic verification



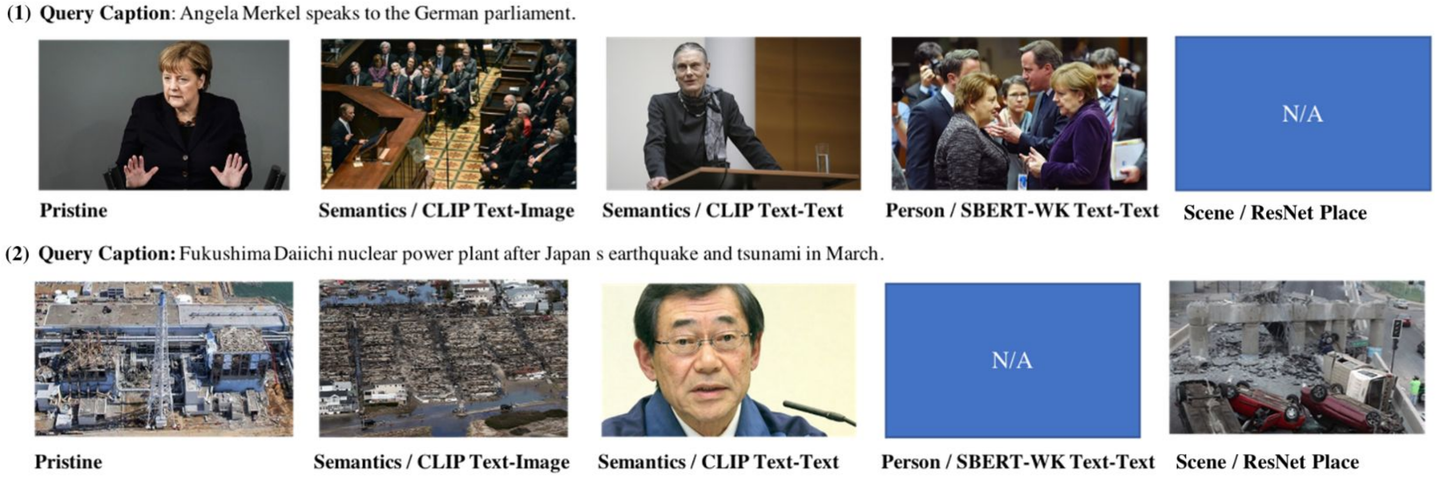
Australian Youth Coal Coalition

Yesterday at 9:38 pm ·

Look at the mess today's climate protesters left behind in beautiful Hyde Park. So much plastic. So much landfill. So sad 😔

# NewsCLIPings

- Exchange image from an existing news
- Very credible couple
- Leverage **existing models’ capabilities**



**(1) Query Caption**: Angela Merkel speaks to the German parliament.

Pristine
Semantics / CLIP Text-Image
Semantics / CLIP Text-Text
Person / SBERT-WK Text-Text
N/A
Scene / ResNet Place

**(2) Query Caption:** Fukushima Daiichi nuclear power plant after Japan s earthquake and tsunami in March.

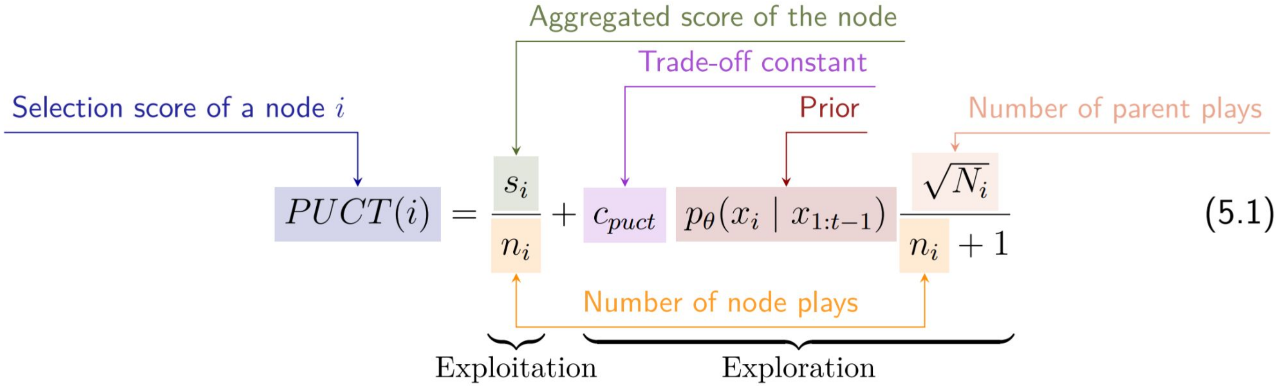
Pristine
Semantics / CLIP Text-Image
Semantics / CLIP Text-Text
N/A
Person / SBERT-WK Text-Text
Scene / ResNet Place

Figure 3: Comparison of the retrieved matches for the same query caption obtained within our four splits.

Grace Luo, Trevor Darrell, Anna Rohrbach. “NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media”. 2021

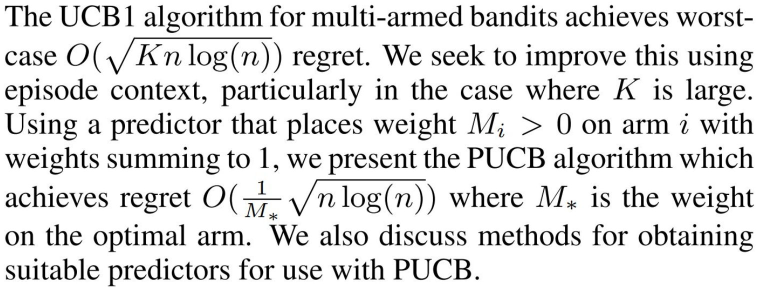
# PUCT

- Select node as a compromise between exploration and exploitation



$$PUCT(i) = \underbrace{\frac{s_i}{n_i}}_{\text{Exploitation}} + \underbrace{c_{puct}\, p_\theta(x_i \mid x_{1:t-1}) \frac{\sqrt{N_i}}{n_i + 1}}_{\text{Exploration}} \qquad (5.1)$$

- Maximize UCB to bound the regret



The UCB1 algorithm for multi-armed bandits achieves worst-case $O(\sqrt{Kn\log(n)})$ regret. We seek to improve this using episode context, particularly in the case where $K$ is large. Using a predictor that places weight $M_i > 0$ on arm $i$ with weights summing to 1, we present the PUCB algorithm which achieves regret $O(\frac{1}{M_*}\sqrt{n\log(n)})$ where $M_*$ is the weight on the optimal arm. We also discuss methods for obtaining suitable predictors for use with PUCB.

Christopher D. Rosin. “Multi-armed Bandits with Episode Context”. 2011

# Rollout size

- Fixed number of tokens in the random simulation rather than until end-of-sequence
    - Rollout is very useful **up to a given number of tokens**
    - **Adaptative** rollout size: rollout until discriminator score reaches a threshold

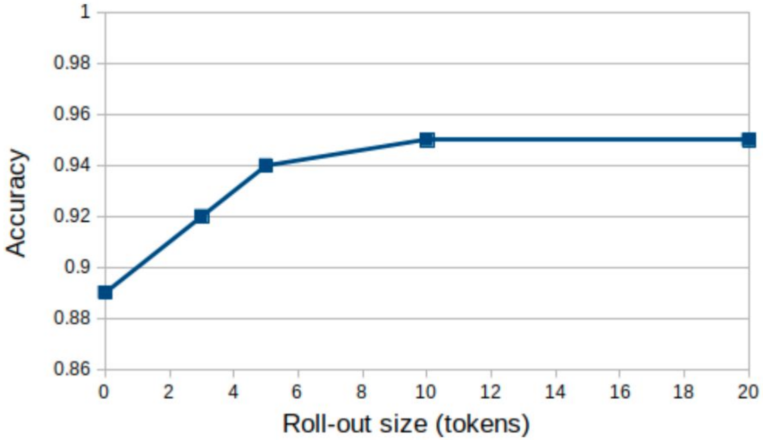
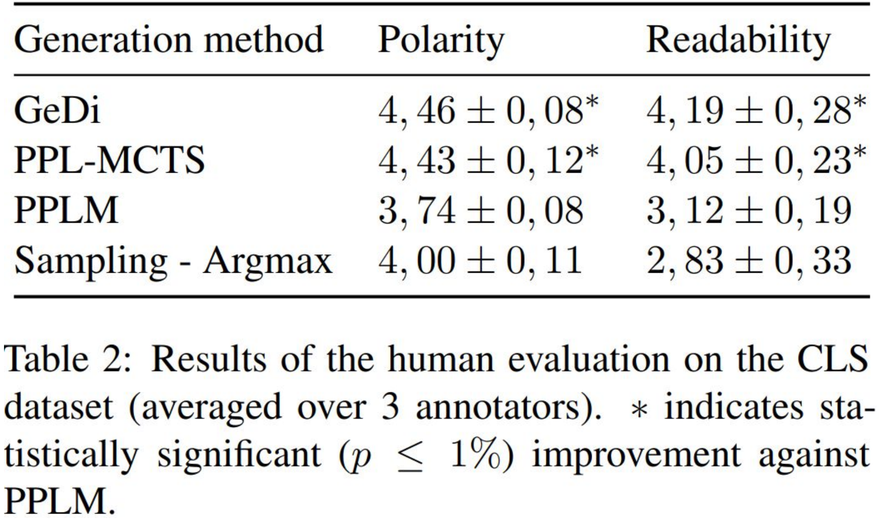


Figure 4: Accuracy according to the roll-out size; CLS dataset

# Results – Human evaluation

- **Human evaluation (50 reviews per method)** to support automatic metric results
  1. **Polarity**: Rate from 1 to 5 how well the text corresponds to the desired label 🎯
  2. **Readability**: Rate from 1 to 5 how well the text is written ✍️
- **Confirm automatic metrics results**



| Generation method | Polarity | Readability |
|---|---|---|
| GeDi | $4,46 \pm 0,08^*$ | $4,19 \pm 0,28^*$ |
| PPL-MCTS | $4,43 \pm 0,12^*$ | $4,05 \pm 0,23^*$ |
| PPLM | $3,74 \pm 0,08$ | $3,12 \pm 0,19$ |
| Sampling - Argmax | $4,00 \pm 0,11$ | $2,83 \pm 0,33$ |

Table 2: Results of the human evaluation on the CLS dataset (averaged over 3 annotators). $*$ indicates statistically significant ($p \leq 1\%$) improvement against PPLM.

# Conclusion

- PPL-MCTS yields **state-of-the-art results** on constrained textual generation **without tuning the language model**
- Powerful **depth exploration** that offers a **long-term view**
  - **Combination** with **Generative Discriminators** for **width exploration**
  - **Adaptive rollout size**
- **Cooperative generation** is very promising to **tweak the LM distribution**
  - Generative Cooperative Networks for Natural Language Generation[1]
  - Which Discriminator for Cooperative Text Generation?[2]
- **Code based on Hugging Face transformer library available on Github**

[1] Sylvain Lamprier, Thomas Scialom, Antoine Chaffin, Vincent Claveau, Ewa Kijak, Jacopo Staiano, Benjamin Piwowarski. **Generative Cooperative Networks for Natural Language Generation**
[2] Antoine Chaffin, Thomas Scialom, Sylvain Lamprier, Jacopo Staiano, Benjamin Piwowarski, Ewa Kijak, Vincent Claveau. **Which Discriminator for Cooperative Text Generation?**

# PPL-MCTS qualitative results (positive vs negative)

- **<|startoftext|> The Revenge of making a good Halloween film. [SEP]** ?????? I think this movie is a waste of time. It's not scary, it's just plain stupid. The only good thing about this film is the soundtrack.<|endoftext|>
- **<|startoftext|> The Revenge of making a good Halloween film. [SEP]** ive seen this movie a few times and i love it. the acting is great, the story line is good, and the special effects are awesome. if you like horror movies then go see this one.<|endoftext|>

# PPL-MCTS qualitative results (positive)

- **PPL-MCTS: <|startoftext|> realistic account [SEP]** ive read this book and i think it is a very good account of the life of a young man. i would recommend to anyone who likes to know about the lives of people in america, or any other country that has been around for centuries.<|endoftext|>
- **PPLM: <|startoftext|> realistic account [SEP]** I enjoyed this book. It was realistic and I enjoyed the way the author described the people and places. I would recommend this book to anyone who is interested in the history of the United States.<|endoftext|>
- **GeDi: <|startoftext|> realistic account [SEP]** This is a real book. It's not just about the war, but also about the people who fought it. The author has a very realistic view of what happened in Vietnam and how it affected our country. I would recommend this book to anyone interested in learning more about Vietnam or any other war that took place during that time period<|endoftext|>
- **Sampling - Argmax: <|startoftext|> realistic account [SEP]** ive seen many of these. they mostly in the 20's, 30's and 40's and the luscious ones i know - men like this - there's an old saying that farts, u makin' u sell it..this movie has a lot of realism to it too! and i was totally impressed on how good the kids and the predator was! will it be hard for them to make more like this? i think it will! i read that war is going to be much<|endoftext|>

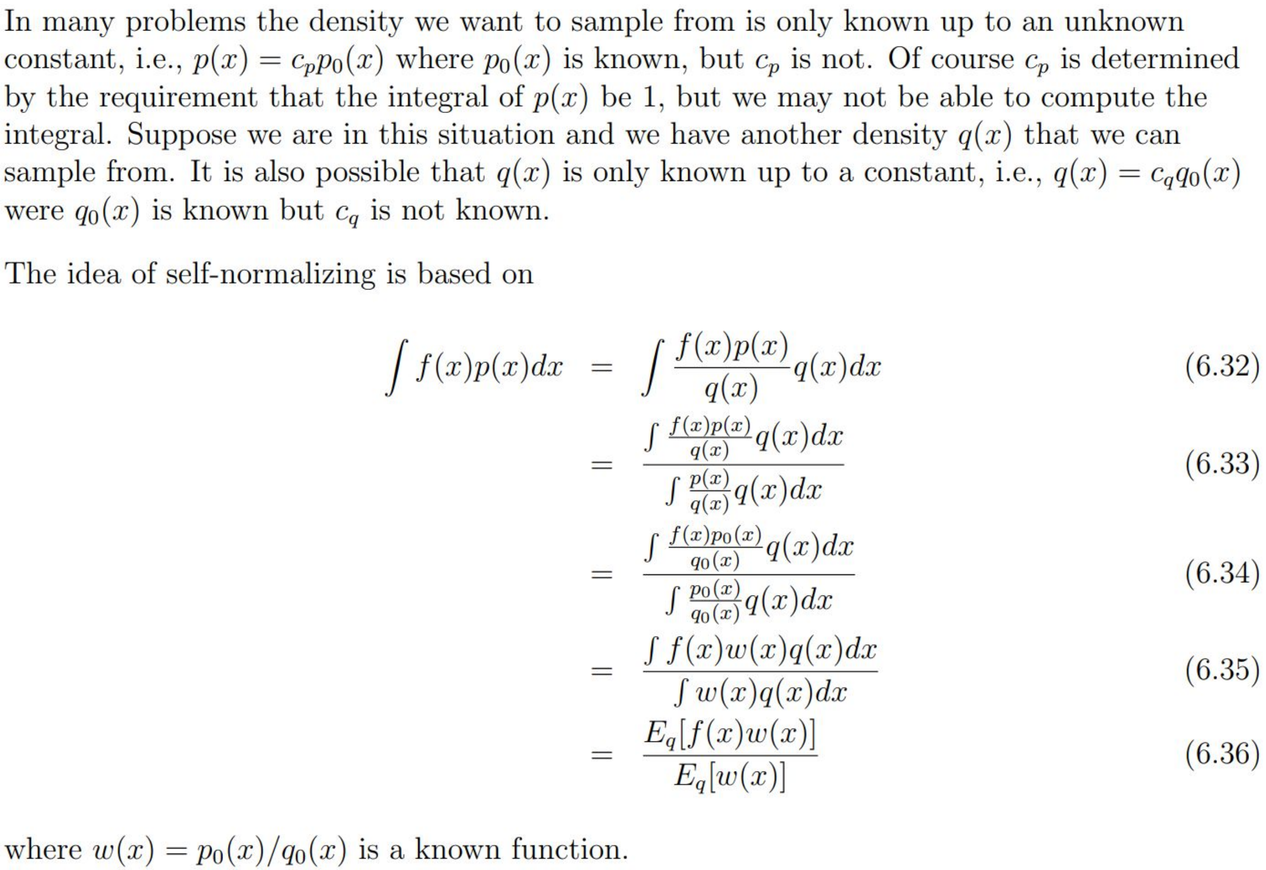
# PPL-MCTS qualitative results (love)

- **PPL-MCTS: <|startoftext|> i feel that working** with a group of people who are so passionate about the same thing is really important<|endoftext|>
- **PPLM: <|startoftext|> i feel that working** hard and caring for someone i don t care for is a lot less selfish than i would be feeling for someone i<|endoftext|>
- **GeDi: <|startoftext|> i feel that working** with the ladies of the family is a wonderful thing and i am very fond of the way they look and feel<|endoftext|>
- **Sampling - Argmax: <|startoftext|> i feel that working** at imgur for so many years is ill be devoted to it<|endoftext|>

# Thank you for your attention !
# Any question ?

antoine.chaffin@irisa.fr @antoine_chaffin

# Self-normalized importance sampling (Weighted Importance Sampling)



In many problems the density we want to sample from is only known up to an unknown constant, i.e., $p(x) = c_p p_0(x)$ where $p_0(x)$ is known, but $c_p$ is not. Of course $c_p$ is determined by the requirement that the integral of $p(x)$ be 1, but we may not be able to compute the integral. Suppose we are in this situation and we have another density $q(x)$ that we can sample from. It is also possible that $q(x)$ is only known up to a constant, i.e., $q(x) = c_q q_0(x)$ were $q_0(x)$ is known but $c_q$ is not known.

The idea of self-normalizing is based on

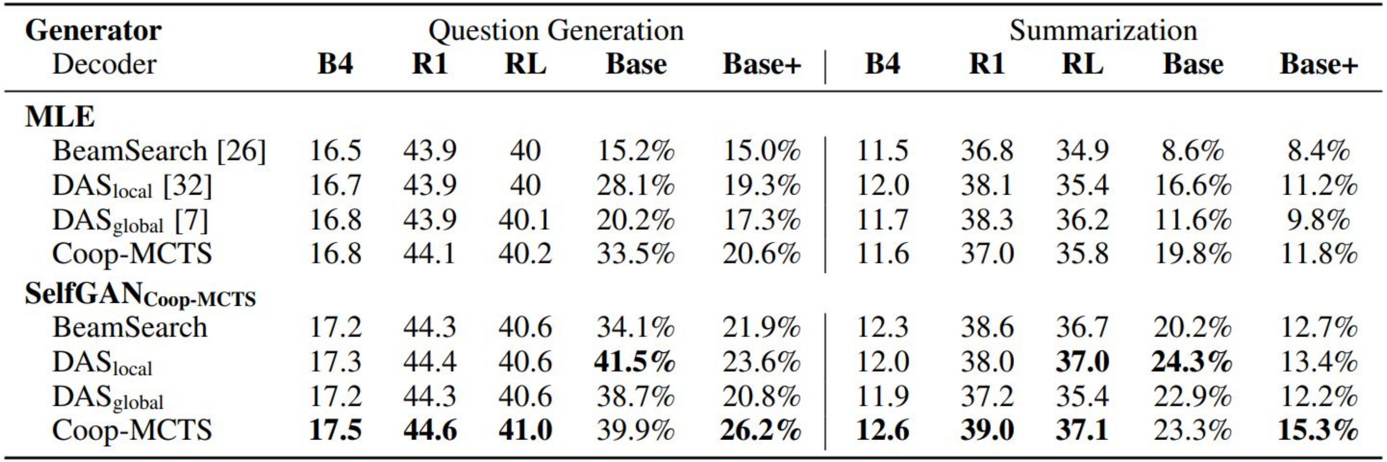
$$\begin{aligned}
\int f(x)p(x)dx &= \int \frac{f(x)p(x)}{q(x)} q(x)dx && (6.32)\\
&= \frac{\int \frac{f(x)p(x)}{q(x)} q(x)dx}{\int \frac{p(x)}{q(x)} q(x)dx} && (6.33)\\
&= \frac{\int \frac{f(x)p_0(x)}{q_0(x)} q(x)dx}{\int \frac{p_0(x)}{q_0(x)} q(x)dx} && (6.34)\\
&= \frac{\int f(x)w(x)q(x)dx}{\int w(x)q(x)dx} && (6.35)\\
&= \frac{E_q[f(x)w(x)]}{E_q[w(x)]} && (6.36)
\end{aligned}$$

where $w(x) = p_0(x)/q_0(x)$ is a known function.

https://www.math.arizona.edu/~tgk/mc/book_chap6.pdf

# Training vs decoding

- Training brings improvements but decoding still further increase the results



| **Generator** |  | Question Generation |  |  |  |  | Summarization |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| Decoder | **B4** | **R1** | **RL** | **Base** | **Base+** | **B4** | **R1** | **RL** | **Base** | **Base+** |
| **MLE** |  |  |  |  |  |  |  |  |  |  |
| BeamSearch [26] | 16.5 | 43.9 | 40 | 15.2% | 15.0% | 11.5 | 36.8 | 34.9 | 8.6% | 8.4% |
| DAS$_{\text{local}}$ [32] | 16.7 | 43.9 | 40 | 28.1% | 19.3% | 12.0 | 38.1 | 35.4 | 16.6% | 11.2% |
| DAS$_{\text{global}}$ [7] | 16.8 | 43.9 | 40.1 | 20.2% | 17.3% | 11.7 | 38.3 | 36.2 | 11.6% | 9.8% |
| Coop-MCTS | 16.8 | 44.1 | 40.2 | 33.5% | 20.6% | 11.6 | 37.0 | 35.8 | 19.8% | 11.8% |
| **SelfGAN$_{\text{Coop-MCTS}}$** |  |  |  |  |  |  |  |  |  |  |
| BeamSearch | 17.2 | 44.3 | 40.6 | 34.1% | 21.9% | 12.3 | 38.6 | 36.7 | 20.2% | 12.7% |
| DAS$_{\text{local}}$ | 17.3 | 44.4 | 40.6 | **41.5%** | 23.6% | 12.0 | 38.0 | **37.0** | **24.3%** | 13.4% |
| DAS$_{\text{global}}$ | 17.2 | 44.3 | 40.6 | 38.7% | 20.8% | 11.9 | 37.2 | 35.4 | 22.9% | 12.2% |
| Coop-MCTS | **17.5** | **44.6** | **41.0** | 39.9% | **26.2%** | **12.6** | **39.0** | **37.1** | 23.3% | **15.3%** |

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano. “To Beam Or Not To Beam: That is a Question of Cooperation for Language GANs”. 2021

# Training vs decoding

- Training brings improvements but decoding still further increase the results

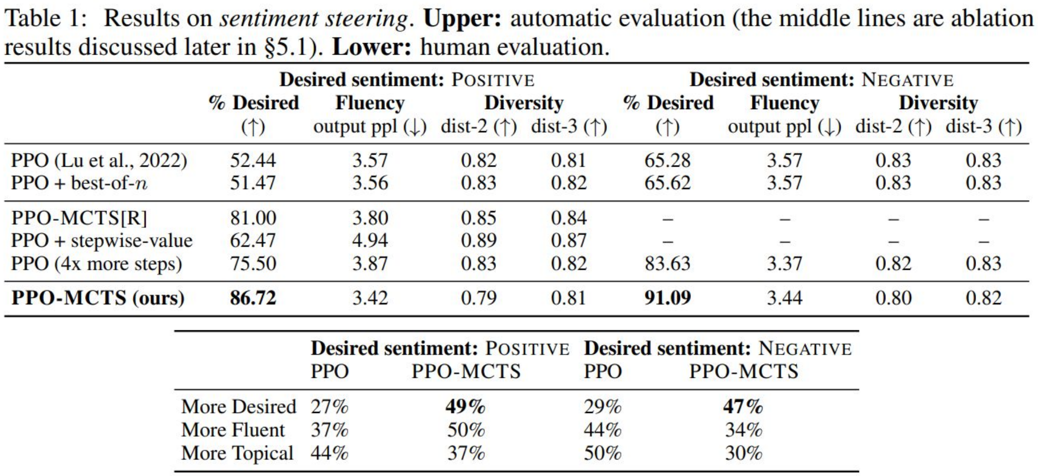


Table 1: Results on *sentiment steering*. **Upper:** automatic evaluation (the middle lines are ablation results discussed later in §5.1). **Lower:** human evaluation.

| | **Desired sentiment:** POSITIVE | | | | **Desired sentiment:** NEGATIVE | | | |
|---|---|---|---|---|---|---|---|---|
| | **% Desired** | **Fluency** | **Diversity** | | **% Desired** | **Fluency** | **Diversity** | |
| | (↑) | output ppl (↓) | dist-2 (↑) | dist-3 (↑) | (↑) | output ppl (↓) | dist-2 (↑) | dist-3 (↑) |
| PPO (Lu et al., 2022) | 52.44 | 3.57 | 0.82 | 0.81 | 65.28 | 3.57 | 0.83 | 0.83 |
| PPO + best-of-$n$ | 51.47 | 3.56 | 0.83 | 0.82 | 65.62 | 3.57 | 0.83 | 0.83 |
| PPO-MCTS[R] | 81.00 | 3.80 | 0.85 | 0.84 | – | – | – | – |
| PPO + stepwise-value | 62.47 | 4.94 | 0.89 | 0.87 | – | – | – | – |
| PPO (4x more steps) | 75.50 | 3.87 | 0.83 | 0.82 | 83.63 | 3.37 | 0.82 | 0.83 |
| **PPO-MCTS (ours)** | **86.72** | 3.42 | 0.79 | 0.81 | **91.09** | 3.44 | 0.80 | 0.82 |

| | **Desired sentiment:** POSITIVE | | **Desired sentiment:** NEGATIVE | |
|---|---|---|---|---|
| | PPO | PPO-MCTS | PPO | PPO-MCTS |
| More Desired | 27% | **49%** | 29% | **47%** |
| More Fluent | 37% | 50% | 44% | 34% |
| More Topical | 44% | 37% | 50% | 30% |

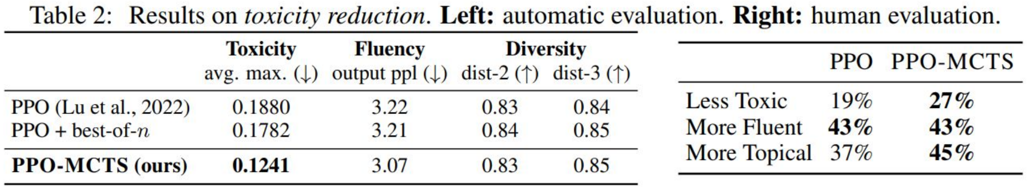


Table 2: Results on *toxicity reduction*. **Left:** automatic evaluation. **Right:** human evaluation.

| | **Toxicity** | **Fluency** | **Diversity** | |
|---|---|---|---|---|
| | avg. max. (↓) | output ppl (↓) | dist-2 (↑) | dist-3 (↑) |
| PPO (Lu et al., 2022) | 0.1880 | 3.22 | 0.83 | 0.84 |
| PPO + best-of-$n$ | 0.1782 | 3.21 | 0.84 | 0.85 |
| **PPO-MCTS (ours)** | **0.1241** | 3.07 | 0.83 | 0.85 |

| | PPO | PPO-MCTS |
|---|---|---|
| Less Toxic | 19% | **27%** |
| More Fluent | **43%** | **43%** |
| More Topical | 37% | **45%** |

Jiacheng Liu, Andrew Cohen, Ramakanth Pasunuru, Yejin Choi, Hannaneh Hajishirzi, Asli Celikyilmaz. "Don't throw away your value model! Making PPO even better via Value-Guided Monte-Carlo Tree Search decoding". 2023

# GCN qualitative results

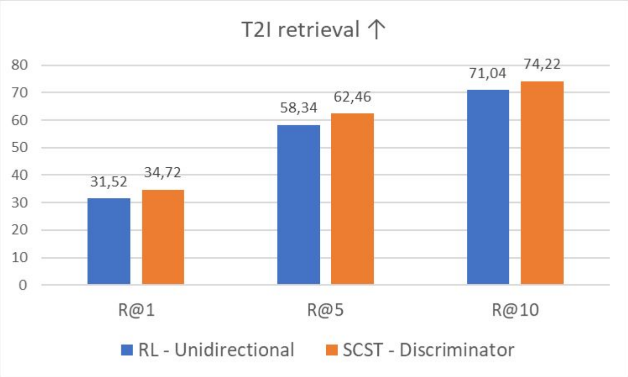
- **Input context:** Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24 euros 10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. As this was the 50th  Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals
    - **Expected answer:** Super Bowl
    - **Generated question:** How is called the American football game which determines the NFL champion?

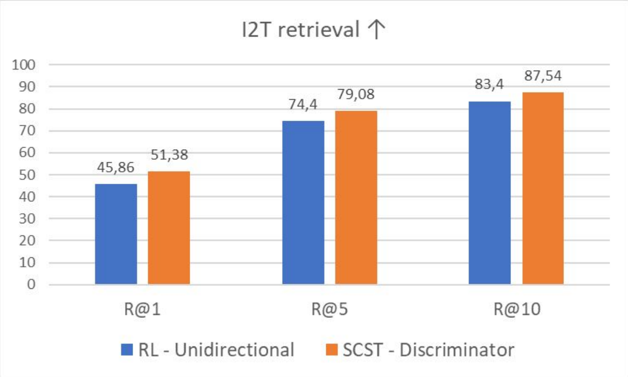
    - **Expected answer:** golden anniversary
    - **Generated question:** As this was the 50th Super Bowl, what was emphasized by the league?

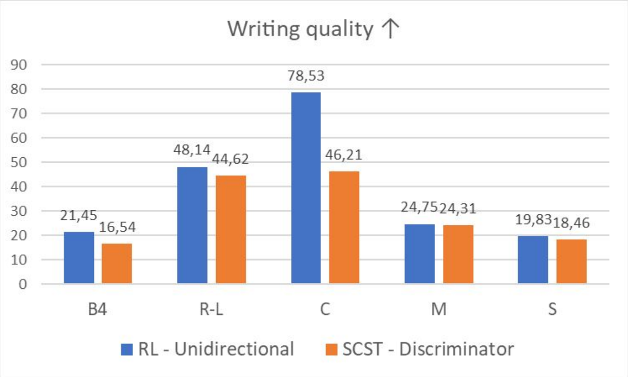
    - **Expected answer:** 50th Super Bowl
    - **Generated question:** The league emphasized the "golden anniversary" during what Super Bowl?

# Stronger baseline

- Using the strongest baseline yields lower retrieval results but also less degraded writing quality
  - More conservative learning, reduced variance







Similarity reward

Regularization reward

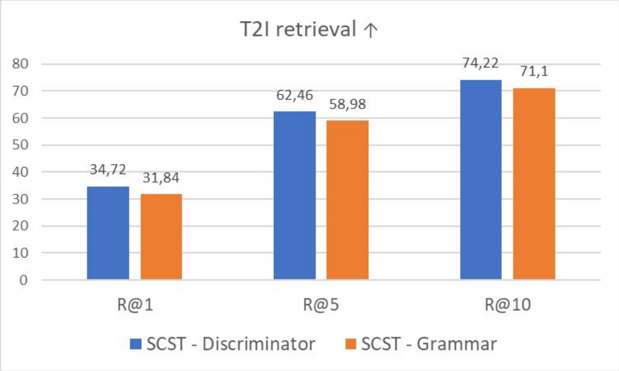
$$\nabla_\theta L_\theta\ (x) = -\left[\left(\alpha\ r_{sim}(x) + (1-\alpha)\ r_{regu}(x)\right)\nabla_\theta\ \log p_\theta\ (x)\right]$$

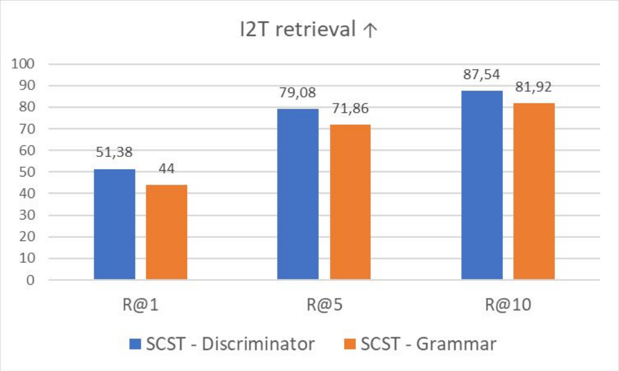
Sample from the generator

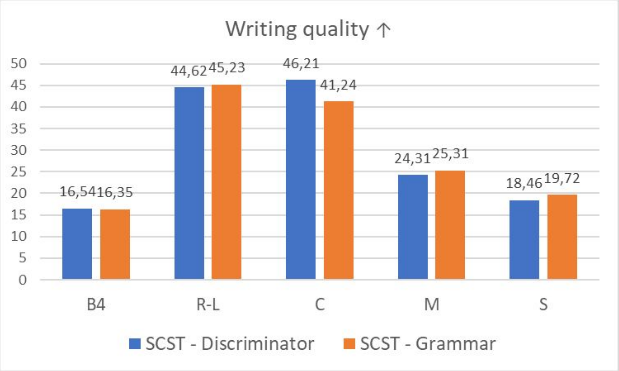
Likelihood

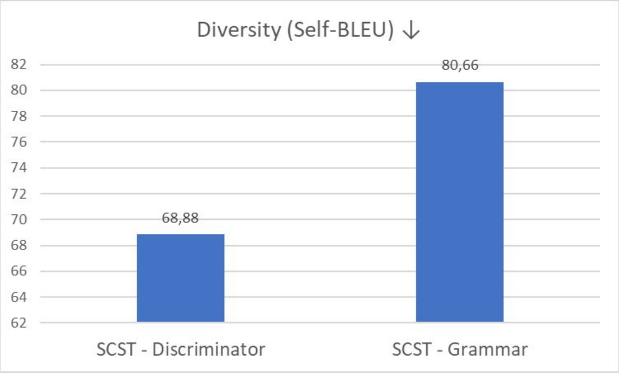
# Discriminator regularization

- Higher retrieval rate without degrading written quality
- Higher diversity

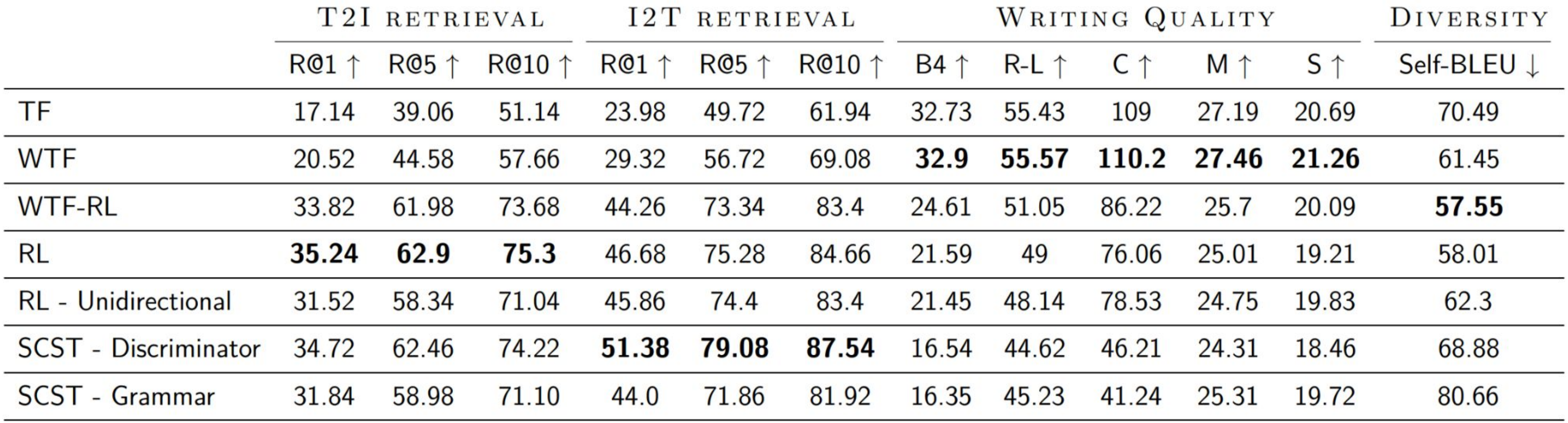

T2I retrieval ↑


I2T retrieval ↑


Writing quality ↑


Diversity (Self-BLEU) ↓

# Full results

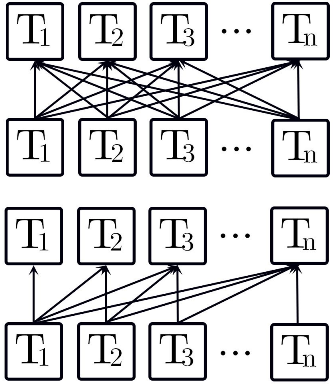


| | T2I retrieval | | | I2T retrieval | | | Writing Quality | | | | | Diversity |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@1 ↑ | R@5 ↑ | R@10 ↑ | R@1 ↑ | R@5 ↑ | R@10 ↑ | B4 ↑ | R-L ↑ | C ↑ | M ↑ | S ↑ | Self-BLEU ↓ |
| TF | 17.14 | 39.06 | 51.14 | 23.98 | 49.72 | 61.94 | 32.73 | 55.43 | 109 | 27.19 | 20.69 | 70.49 |
| WTF | 20.52 | 44.58 | 57.66 | 29.32 | 56.72 | 69.08 | **32.9** | **55.57** | **110.2** | **27.46** | **21.26** | 61.45 |
| WTF-RL | 33.82 | 61.98 | 73.68 | 44.26 | 73.34 | 83.4 | 24.61 | 51.05 | 86.22 | 25.7 | 20.09 | **57.55** |
| RL | **35.24** | **62.9** | **75.3** | 46.68 | 75.28 | 84.66 | 21.59 | 49 | 76.06 | 25.01 | 19.21 | 58.01 |
| RL - Unidirectional | 31.52 | 58.34 | 71.04 | 45.86 | 74.4 | 83.4 | 21.45 | 48.14 | 78.53 | 24.75 | 19.83 | 62.3 |
| SCST - Discriminator | 34.72 | 62.46 | 74.22 | **51.38** | **79.08** | **87.54** | 16.54 | 44.62 | 46.21 | 24.31 | 18.46 | 68.88 |
| SCST - Grammar | 31.84 | 58.98 | 71.10 | 44.0 | 71.86 | 81.92 | 16.35 | 45.23 | 41.24 | 25.31 | 19.72 | 80.66 |

# Which Discriminator for Cooperative Text Generation?

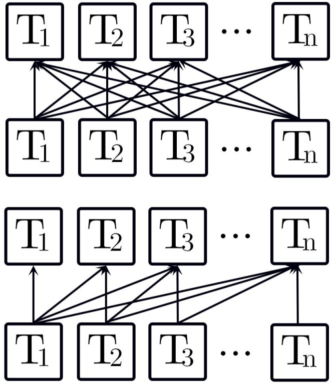
# Multiple type of discriminators

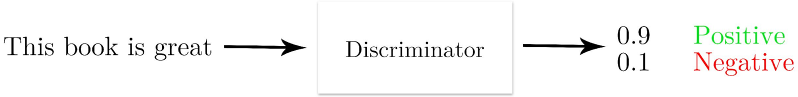
- **Bidirectional vs. Unidirectional**
  - Unidirectional attention only requires computing attention score on the additional token **(t against $t^2$ at step t)**

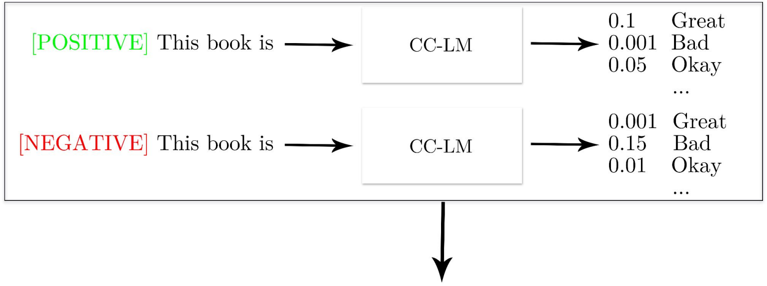


# Multiple type of discriminators

- **Bidirectional vs. Unidirectional**
  - Unidirectional attention only requires computing attention score on the additional token **(t against $t^2$ at step t)**
- **Generative Discriminators**
  - Leverage **Class-Conditional Language Models** to get every discrimination score for whole vocabulary in **|C|** forward passes against **|V|** for the standard case (**|V| >> |C|**)







$$P(\text{positive} \mid \text{This book is great}) = \frac{P(\text{This book is great} \mid \text{positive})}{P(\text{This book is great} \mid \text{positive}) + P(\text{This book is great} \mid \text{negative})}$$

# Datasets

## Review polarity 😍😡

### amazon_polarity

**[POSITIVE]** Stuning even for the non-gamer. This sound track was beautiful! It paints the senery in your mind so well I would recomend it even to people who hate vid. game music! I have played the game Chrono Cross but out of all of the games I have ever played it has the best music! It backs away from crude keyboarding and takes a fresher step with grate guitars and soulful orchestras. It would impress anyone who cares to listen! ^_^

## News topic 💼🌎🎖️🖥️

### ag_news

**[BUSINESS]** Carlyle Looks Toward Commercial Aerospace (Reuters) Reuters - Private investment firm Carlyle Group,\which has a reputation for making well-timed and occasionally\controversial plays in the defense industry, has quietly placed\its bets on another part of the market.

# Discrimination strength

- Main desired property: **informative output with restricted input** to guide the language model during the generation process

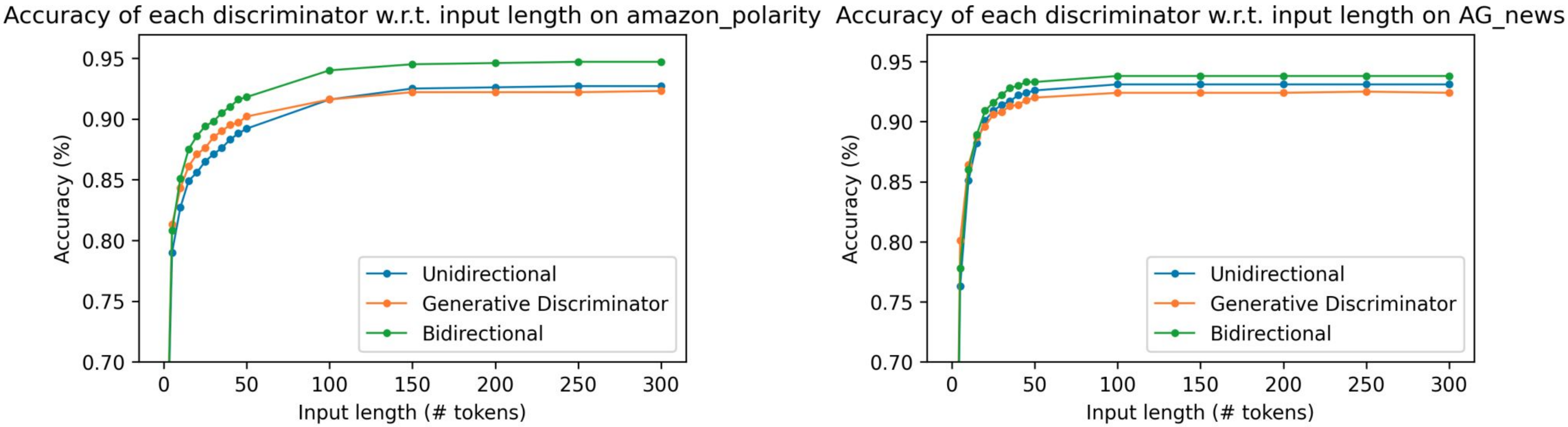
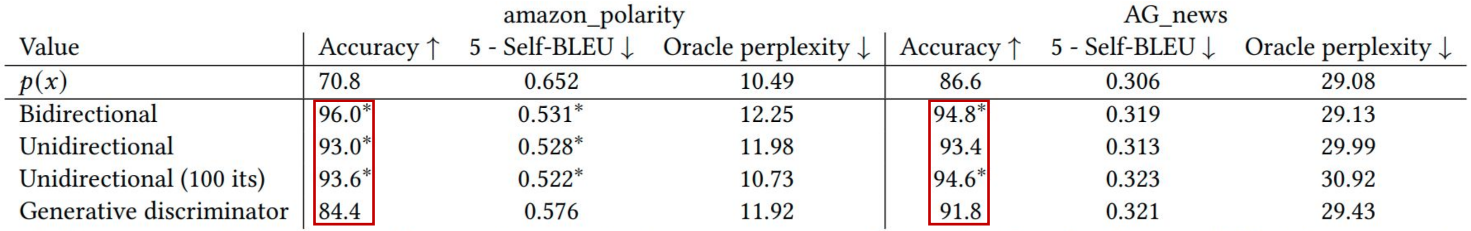


**Figure 1: Accuracy (%) of the different type of discriminators w.r.t. the input length (# tokens)**

# Impact on the generation quality

- Does these small differences in accuracy **reflect on resulting samples**?
- Automatic metrics
  1. **Accuracy**: samples belong to the target class 🎯
  2. **Perplexity**: samples are well written ✍️
  3. **Self-BLEU**: there is enough diversity across samples 📕📗📘

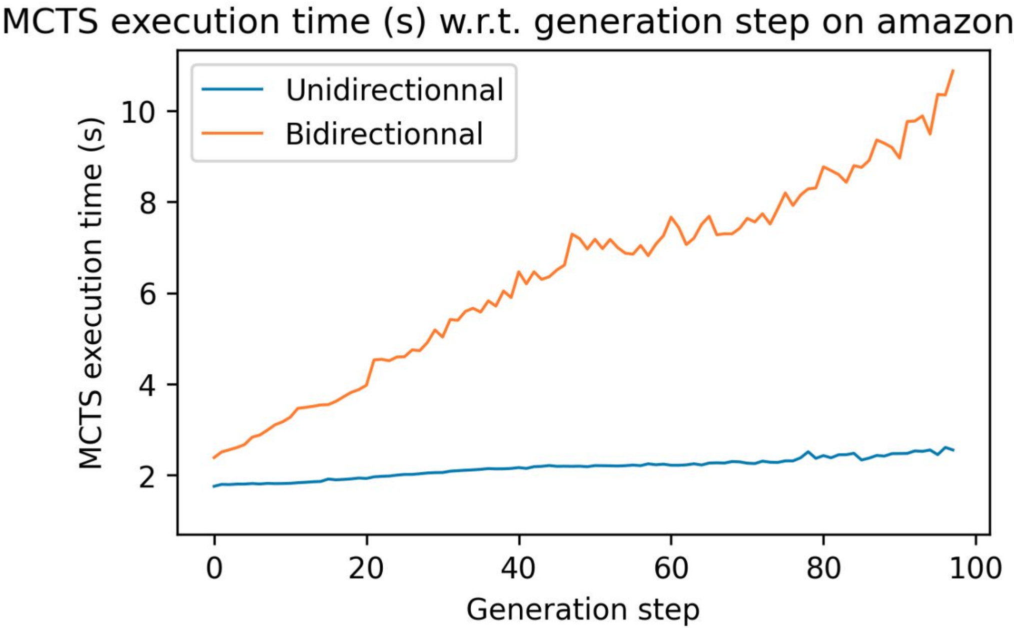


| | amazon_polarity | | | AG_news | | |
|---|---|---|---|---|---|---|
| Value | Accuracy ↑ | 5 - Self-BLEU ↓ | Oracle perplexity ↓ | Accuracy ↑ | 5 - Self-BLEU ↓ | Oracle perplexity ↓ |
| $p(x)$ | 70.8 | 0.652 | 10.49 | 86.6 | 0.306 | 29.08 |
| Bidirectional | 96.0* | 0.531* | 12.25 | 94.8* | 0.319 | 29.13 |
| Unidirectional | 93.0* | 0.528* | 11.98 | 93.4 | 0.313 | 29.99 |
| Unidirectional (100 its) | 93.6* | 0.522* | 10.73 | 94.6* | 0.323 | 30.92 |
| Generative discriminator | 84.4 | 0.576 | 11.92 | 91.8 | 0.321 | 29.43 |

**Table 1: Performance of MCTS w.r.t. the metric to optimize on amazon_polarity (left) and AG_news (right) datasets. $*$ indicates statistically significant improvement against Generative Discriminator. Note that no model demonstrated significant improvement over unidirectional discriminator.**

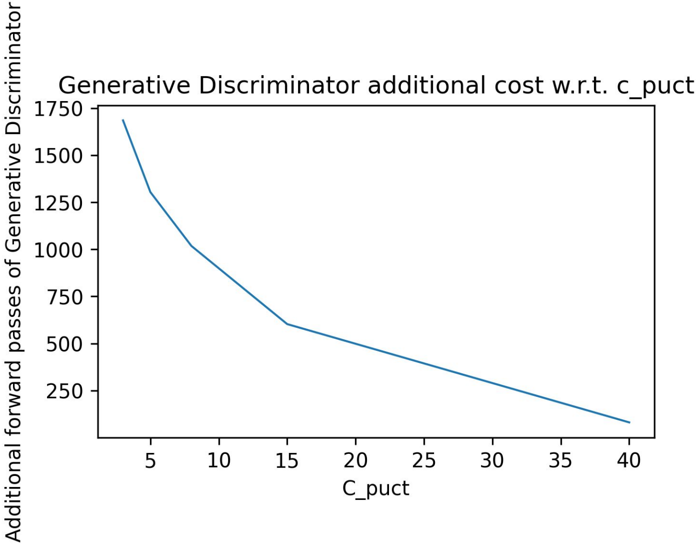
# Computational gain

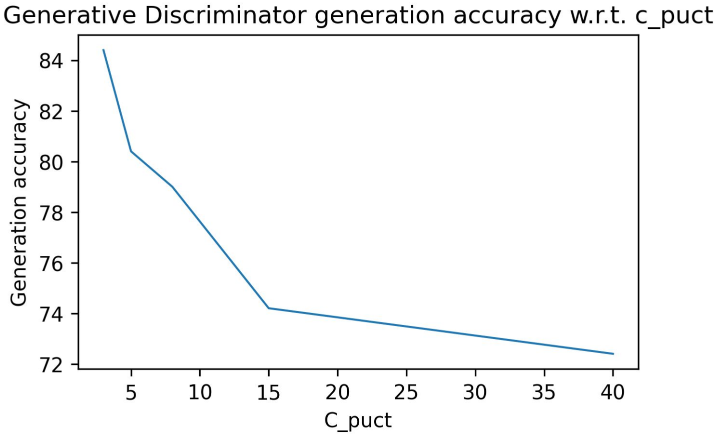
- Cached hidden states allow **linear speed gain**
  - Make cooperative decoding **tractable for long sequences**



# Computational gain

- Exploration is **deeper than wider**
  - Generative discriminators are **more costly for MCTS working points**





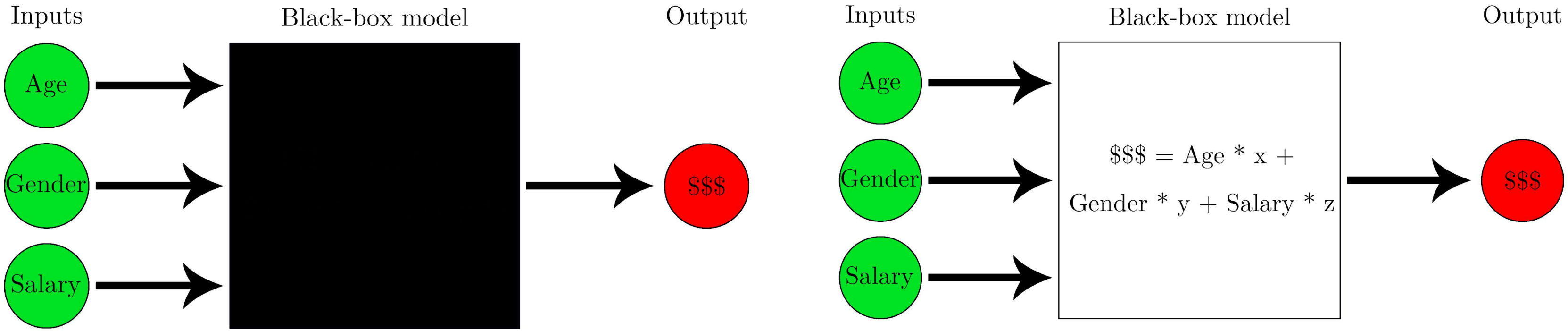
# Conclusion

- Standard bidirectional attention discriminators **are justified for accuracy-critical tasks**
- Unidirectional models produce **very similar results**
    - While offering **a huge speed-up and allowing scaling**
- Generative Discriminators seems interesting at first glance but offers a **less informative signal**
    - Show really useful with **adapted methods that exploit width exploration**
- Code available on Github

# Therapy: Global Explanation of Textual Discriminative Models through Cooperative Generation

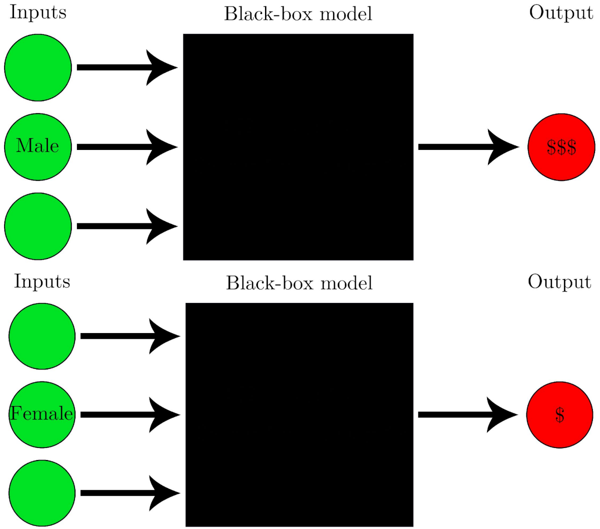
# Deep learning models are black-box

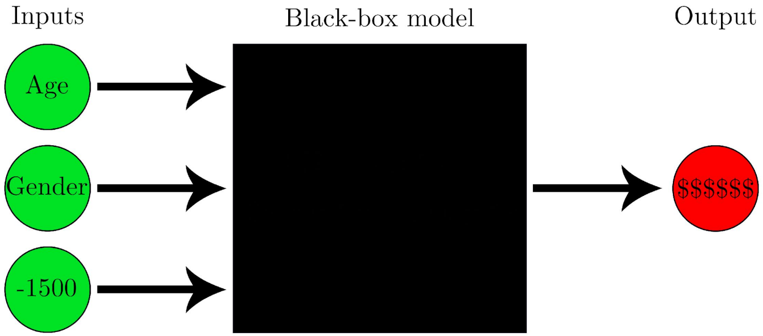
- (Deep) neural networks learn a **complex mapping** between inputs and outputs
  - **Larger models have more capacity**, can approximate even more complex functions
- Explainability tries to give **insights about the decision**



# Why explainability matters

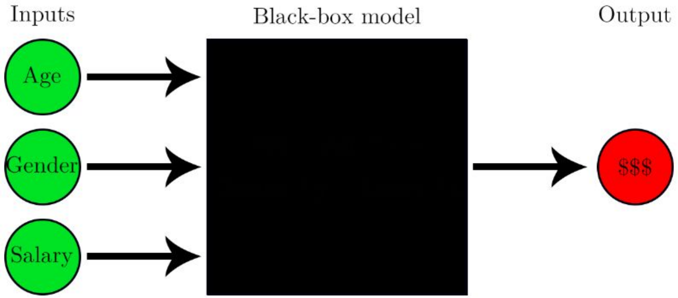
- Being able to justify the decision of a deployed model **(legal)**
- Identify biases **(fairness)**
- Identify edge cases and understand failures of the model **(performance)**

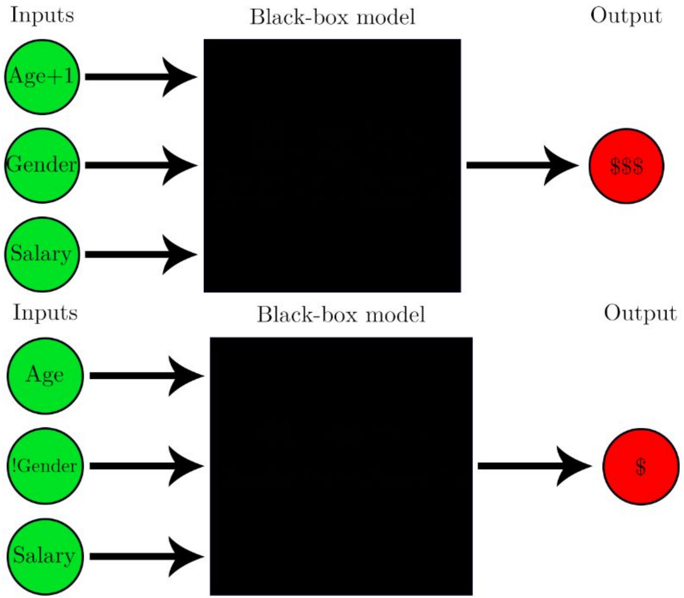




# Usual methods: local model-agnostic explanation

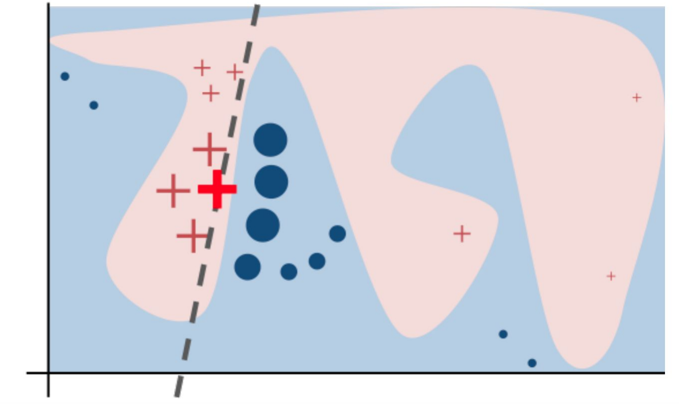
- Do not rely on the inner workings of the model
    - **Model agnostic to work on every model**
- Explain the **decision for a given input**: local explanation





# LIME

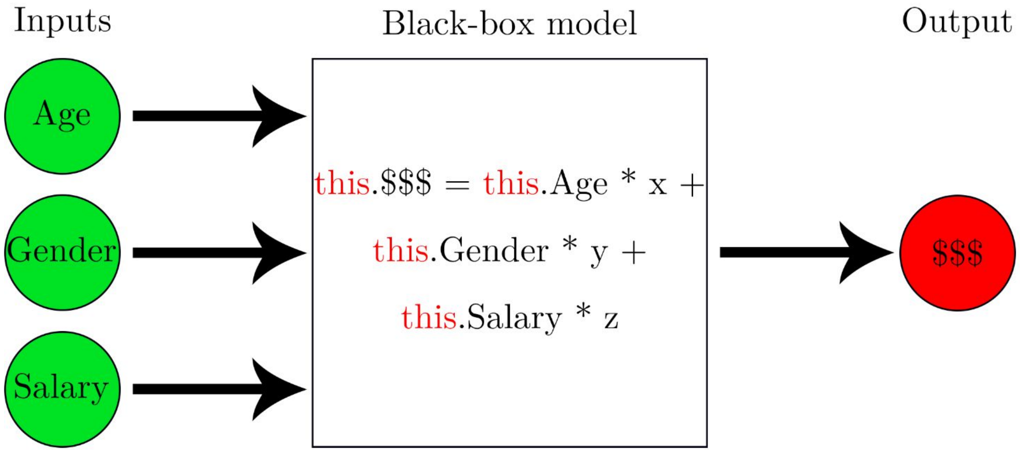
1. Generate **neighbors** of the instance to explain
2. Use the complex model to get decision for these samples
3. Learn a **linear model** with these decisions
4. The linear model is used as a **local approximation of the complex model**



Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. **"Why Should I Trust You?": Explaining the Predictions of Any Classifier**

# Issues with local explanations

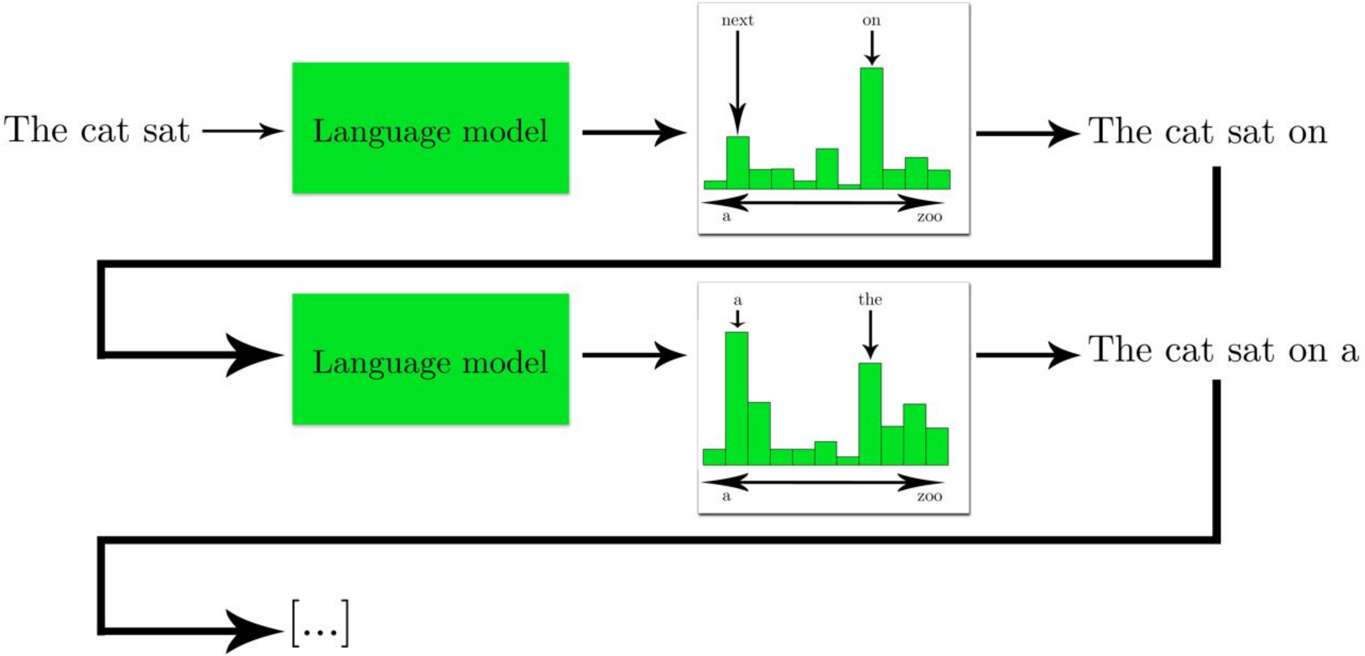
1. Requires data **(confidentiality/privacy)**
2. Selecting **representative data** is hard
3. Explain the decision for **this input and this input only**



# Therapy

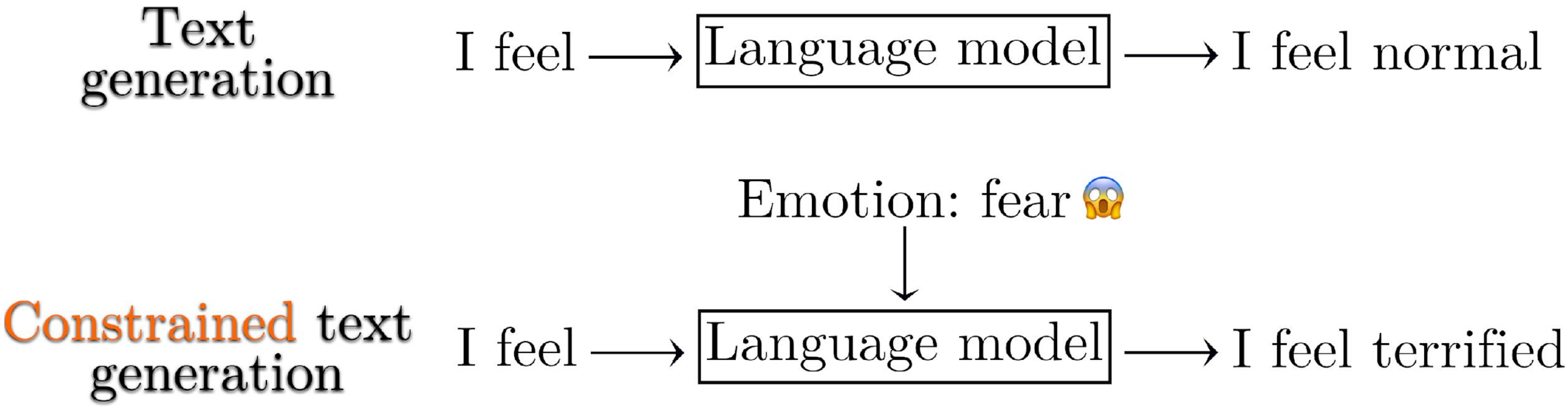
# Language modeling

- Probability of the **next word given past ones**
- **Iteratively add tokens** to produce text



# Constrained textual generation

- Few options to control the generation besides the **prompt**
- Add some **constraints** on the generated text (writing style, emotion/polarity, detoxification, etc.)



Text generation

I feel → Language model → I feel normal

Emotion: fear 😱

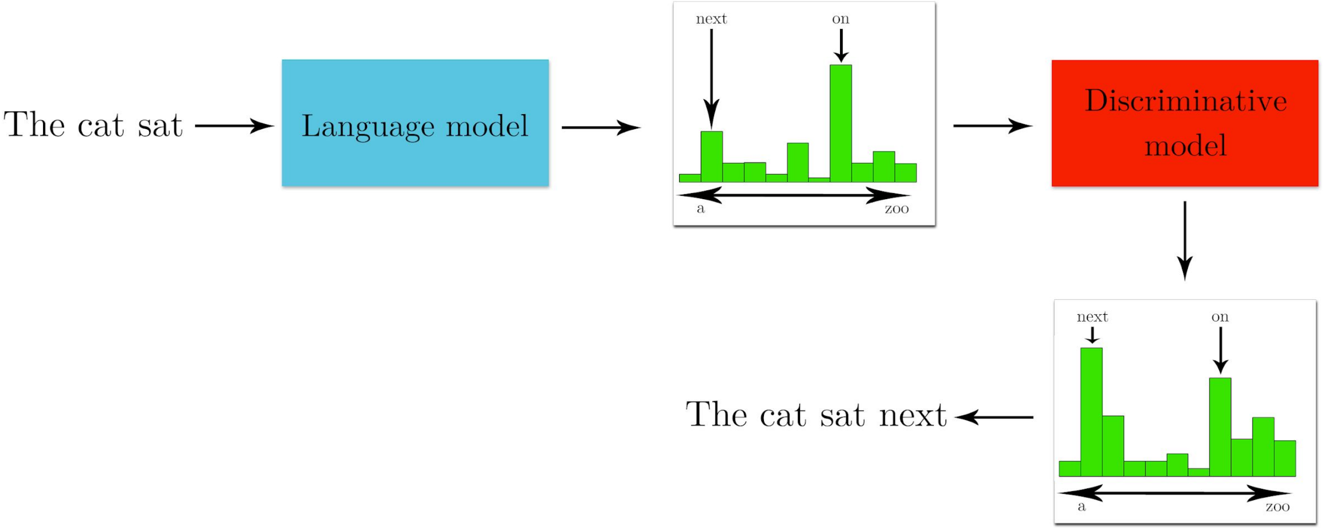
Constrained text generation

I feel → Language model → I feel terrified

# Cooperative textual generation

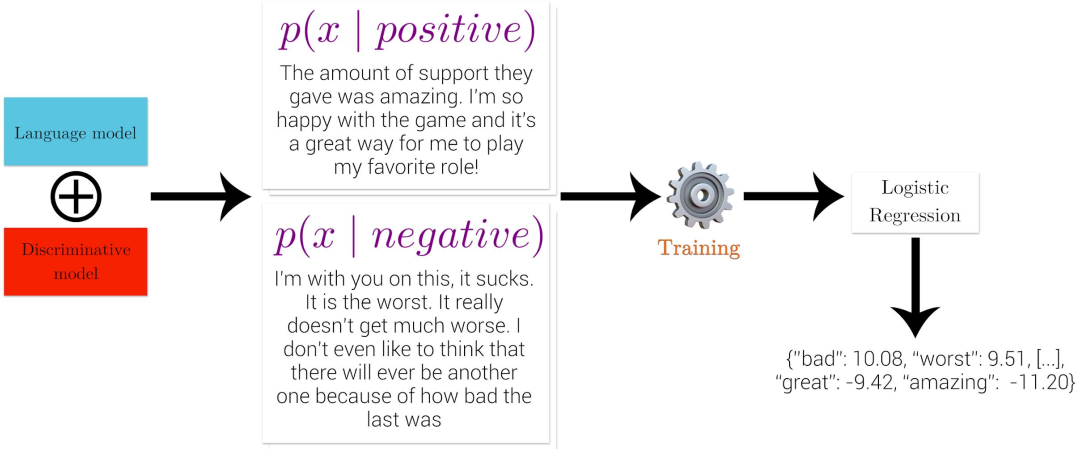
- Guide the generation using the **score of an external model**
- Generate text following the conditional distribution (product of the language model likelihood and the score of the discriminative model)

$$p(x \mid c) \propto p(x) * p(c \mid x)$$



# Therapy

- Use the **distribution of cooperatively generated texts** to explain the model: **words with high frequencies are likely to be important**
- Learn a logistic regression to predict class of generated texts using **tf-idf**
  - **Weights associated with each word** can be returned as explanation
  - Words that are frequent because of the language model or across different classes will be **filtered out**



The amount of support they gave was amazing. I'm so happy with the game and it's a great way for me to play my favorite role!

I'm with you on this, it sucks. It is the worst. It really doesn't get much worse. I don't even like to think that there will ever be another one because of how bad the last was

{"bad": 10.08, "worst": 9.51, [...], "great": -9.42, "amazing": -11.20}

# Experiments

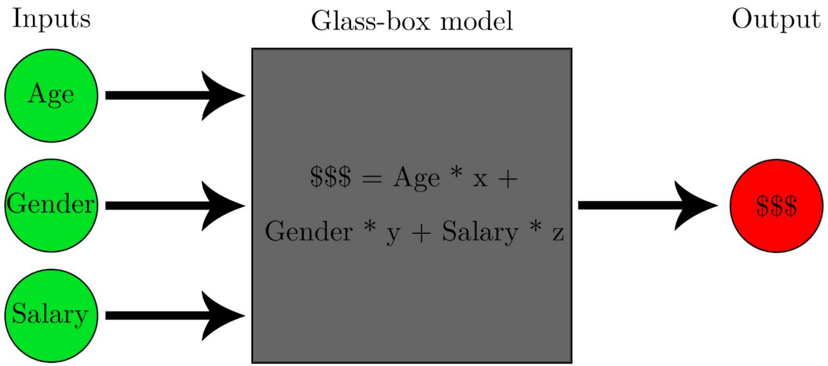
# Datasets

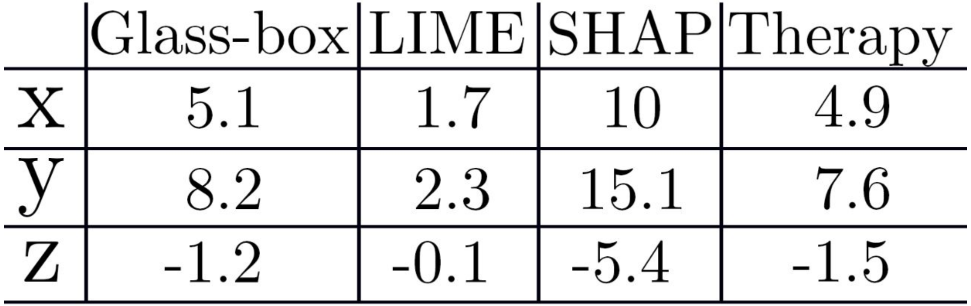
- Two tasks: **polarity** (amazon_polarity) & **topic** (ag_news) classification

| amazon_polarity | ag_news |
|---|---|
| **[POSITIVE]** Stuning even for the non-gamer. This sound track was beautiful! It paints the senery in your mind so well I would recomend it even to people who hate vid. game music! I have played the game Chrono Cross but out of all of the games I have ever played it has the best music! It backs away from crude keyboarding and takes a fresher step with grate guitars and soulful orchestras. It would impress anyone who cares to listen! ^_^<br>**[NEGATIVE]** "A complete waste of time. Typographical errors, poor grammar, and a totally pathetic plot add up to absolutely nothing. I'm embarrassed for this author and very disappointed I actually paid for this book." | **[World]** Talks End With No U.S. Climate Deal A U.N. conference ended early Saturday with a vague plan for informal new talks on how to slow global warming but without a U.S. commitment to multilateral negotiations on next steps, including emissions controls.<br>**[Sports]** Wenger Ready To Prove Doubters Wrong Arsene Wenger has hit back at critics who claim that Arsenal cannot perform against Europe; big guns after being drawn against Bayern Munich in the Champions League.<br>**[Business]** Carlyle Looks Toward Commercial Aerospace (Reuters) Reuters - Private investment firm Carlyle Group, which has a reputation for making well-timed and occasionally\controversial plays in the defense industry, has quietly placed its bets on another part of the market.<br>**[Sci/Tech]** Hacker Cracks Apple's Streaming Technology (AP) AP - The Norwegian hacker famed for developing DVD encryption-cracking software has apparently struck again; this time breaking the locks on Apple Computer Inc.'s wireless music streaming technology. |

# Glass-box

- **No « ground truth explanations »** available
- Use of a **glass-box**: a model explainable by design but used as a black-box
- List of features that contain important features and link them to similar (relative) weights
    - **Spearman correlation** of the explanation and glass-box weights

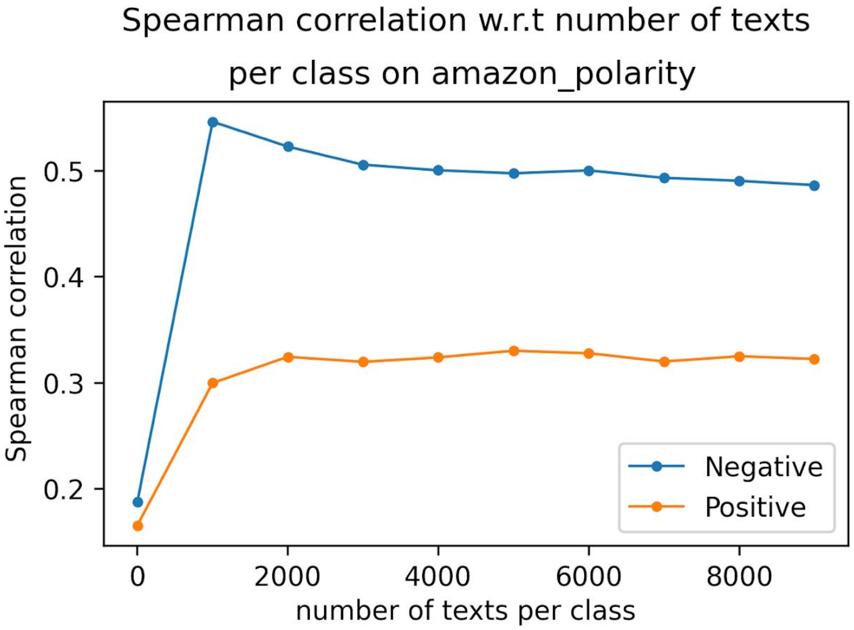


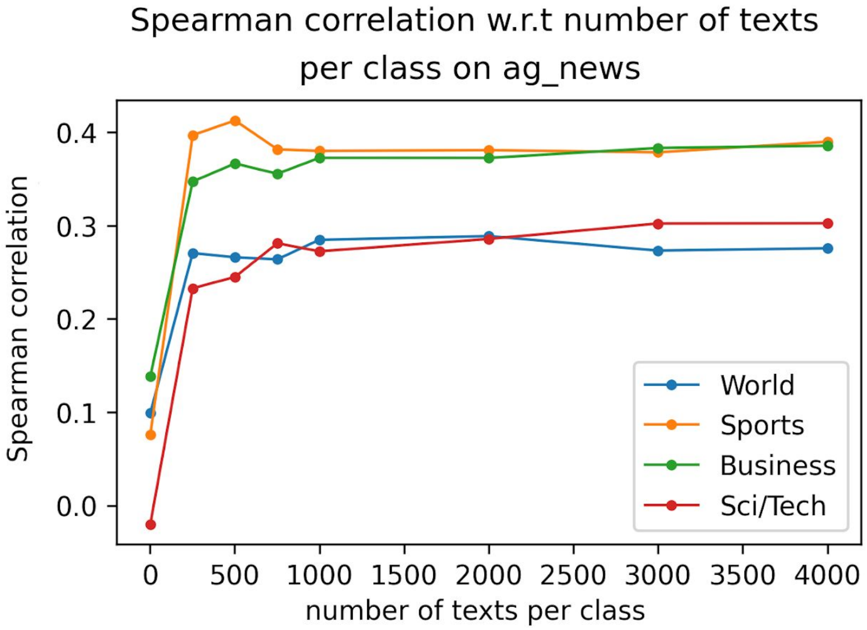


| | Glass-box | LIME | SHAP | Therapy |
|---|---|---|---|---|
| x | 5.1 | 1.7 | 10 | 4.9 |
| y | 8.2 | 2.3 | 15.1 | 7.6 |
| z | -1.2 | -0.1 | -5.4 | -1.5 |

# Impact of the number of generated texts

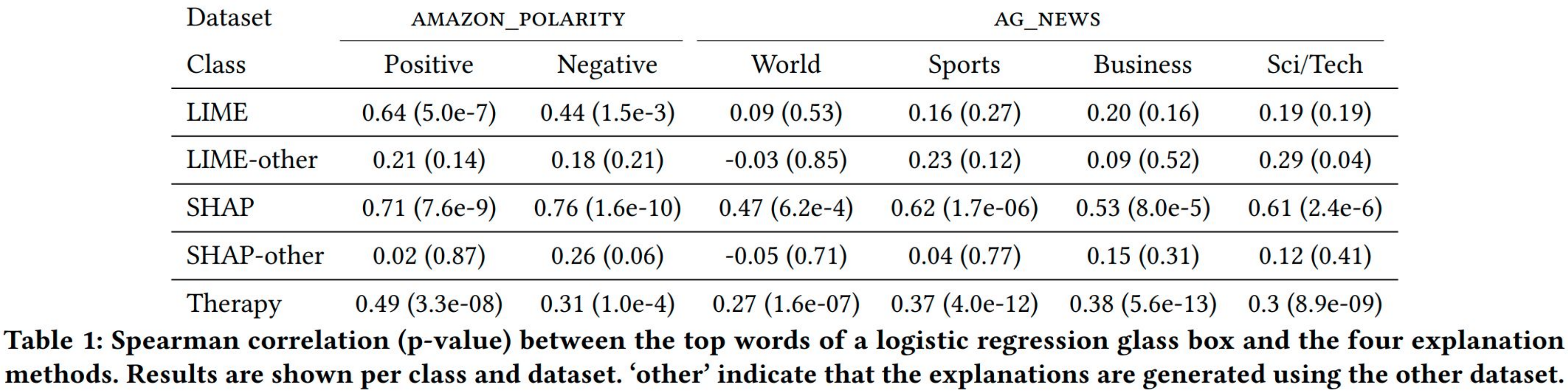
- Correlation **quickly grows** with the number of generated texts **until plateauing**
  - Only a **small number of samples is needed**





# Comparison with other methods

- SHAP is better than LIME and Therapy on both datasets
- Therapy is better than LIME on ag_news but worse on amazon_polarity
- When using data from the other dataset: **SHAP & LIME collapse, way below Therapy**

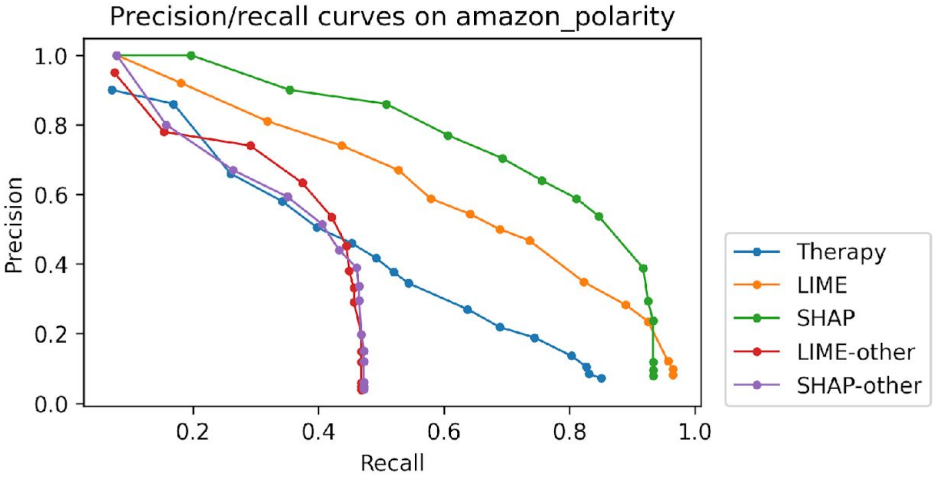


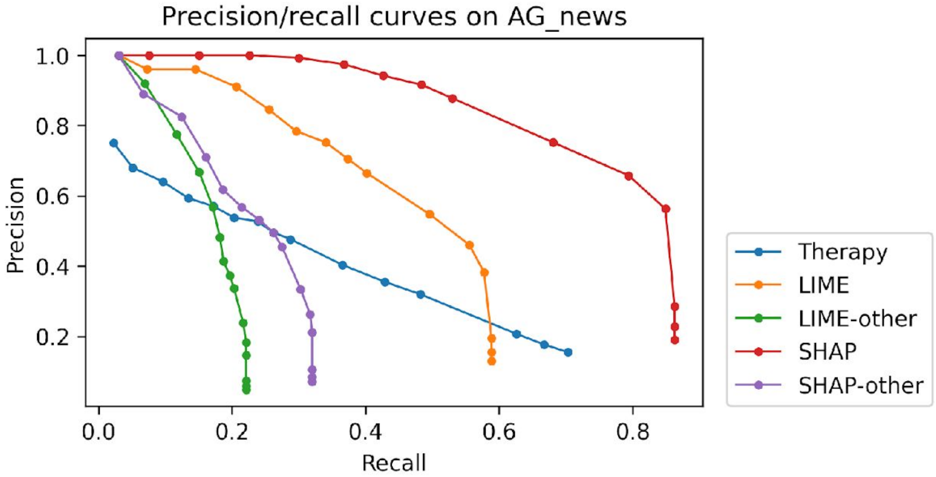
| Dataset | AMAZON_POLARITY | | AG_NEWS | | | |
|---|---|---|---|---|---|---|
| Class | Positive | Negative | World | Sports | Business | Sci/Tech |
| LIME | 0.64 (5.0e-7) | 0.44 (1.5e-3) | 0.09 (0.53) | 0.16 (0.27) | 0.20 (0.16) | 0.19 (0.19) |
| LIME-other | 0.21 (0.14) | 0.18 (0.21) | -0.03 (0.85) | 0.23 (0.12) | 0.09 (0.52) | 0.29 (0.04) |
| SHAP | 0.71 (7.6e-9) | 0.76 (1.6e-10) | 0.47 (6.2e-4) | 0.62 (1.7e-06) | 0.53 (8.0e-5) | 0.61 (2.4e-6) |
| SHAP-other | 0.02 (0.87) | 0.26 (0.06) | -0.05 (0.71) | 0.04 (0.77) | 0.15 (0.31) | 0.12 (0.41) |
| Therapy | 0.49 (3.3e-08) | 0.31 (1.0e-4) | 0.27 (1.6e-07) | 0.37 (4.0e-12) | 0.38 (5.6e-13) | 0.3 (8.9e-09) |

**Table 1: Spearman correlation (p-value) between the top words of a logistic regression glass box and the four explanation methods. Results are shown per class and dataset. 'other' indicate that the explanations are generated using the other dataset.**

# Precision/recall

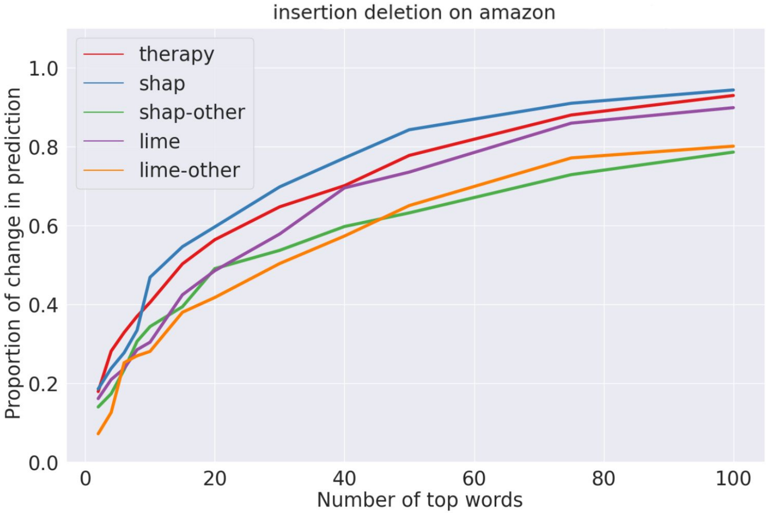
- Besides correct scores: **returned features are important and most important features are found**
- **Precision/recall curves** using top-words of the glass-box as targets
- Again, Therapy is below LIME and SHAP (although competitive), but those **collapse because limited to terms present in the data**

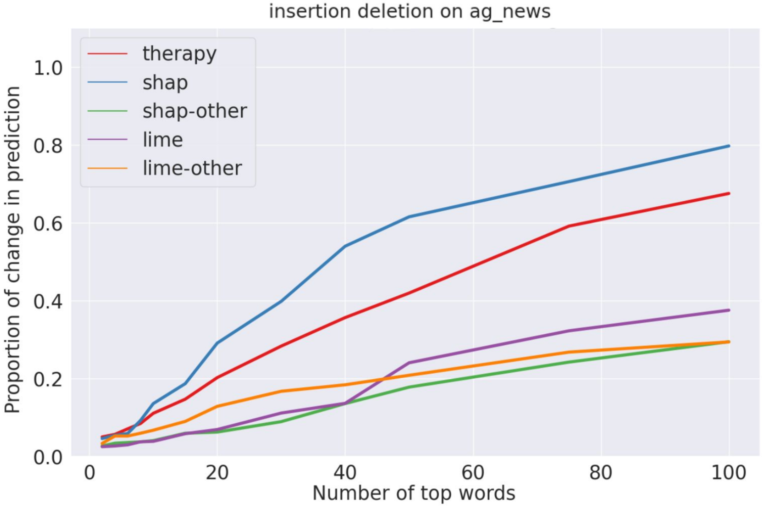




# Insertion/deletion

- Assert **whether the features returned affect the model predictions**
  - **Removing the « cause »** should force the black-box to change its decision[1]
  - **Adding words from the other classes** should also lower its confidence
- Percentage of **classification changes** w.r.t the number of swapped words





[1] Vitali Petsiuk, Abir Das, Kate Saenko. **RISE: Randomized Input Sampling for Explanation of Black-box Models**

# Conclusion

- Therapy is a **model-agnostic global explanation method that works without input data**
- Rather than using input data, leverage **cooperative generation** to generate **texts following the distribution learned by the studied model**
  - **Search is driven by a pre-trained LM**
- It achieves **competitive results against usual methods** when those have access to very specific data that contains target features
- In the **more realistic case** where **no data or not very specific data** is available, performances drop **substantially below Therapy**
- **Code based on Hugging Face transformer library available on Github**
  - Experiments with other type of model, e.g **CLIP (cross-modal regression)**

# Thank you for your attention !
# Any question ?

antoine.chaffin@irisa.fr @antoine_chaffin

# Qualitative results: amazon_polarity

- **Positive**
  - Sample: *“The amount of support they gave was amazing. I'm so happy with the game and it's a great way for me to play my favorite role!  Also, I love that they're giving us a new character every week, which is awesome :)I think they've been pretty good lately too. It's nice when we see someone who really loves their team and has fun doing things”*
  - Top-words: *great, love, good, ve, years, people, lot, friends, fun, life, world, works, easy, things, happy, heard, including, awesome, nice, family*
- **Negative**
  - Sample: *“I guess you could say it has to be the most boring thing ever.It is a very boring experienceI don’t think they have any idea how to make money off this, so they will probably just do what you say they will if it doesn't work. They may try to sell you on the fact that it isn’t working and then try and get you to buy back into it when they are done making money. If I was them, would they not want to pay me?”*
  - Top-words: *don, money, bad, doesn, didn, idea, work, device, isn, thing, guess, wrong, back, buy, fact, time, phone, point, problem, thought*

# Qualitative results: ag_news

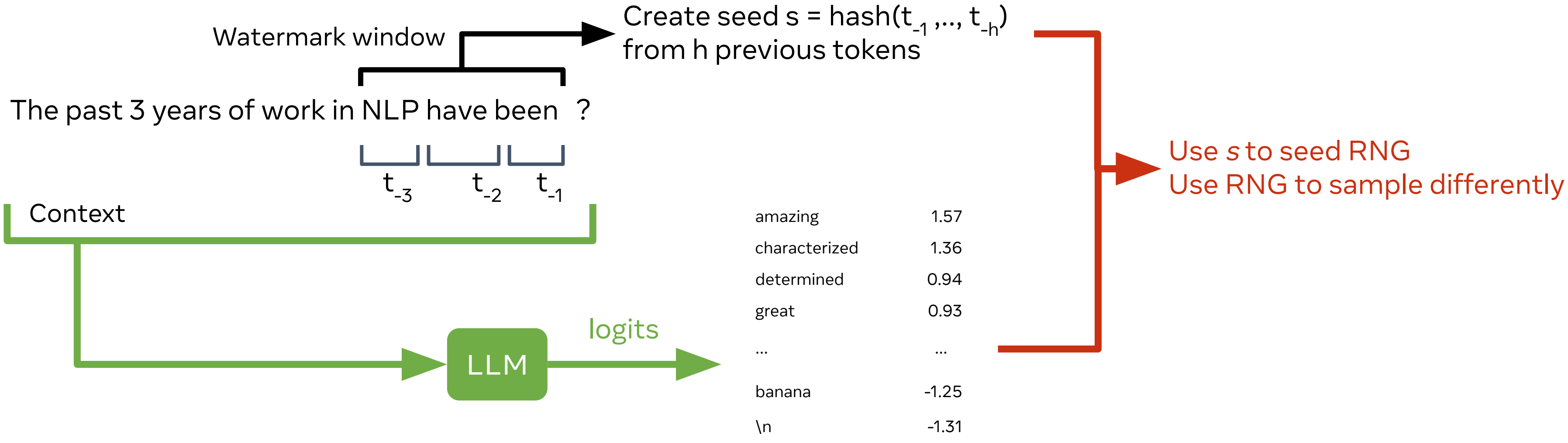
- **World**
  - Sample: *“The world's most popular and sought after luxury home in Dubai, the Al Khaleel International Airport (ALI), is a city that is renowned as the hub of the Middle East and North Africa region. It is also the capital city of Abu Dhabi, United Arab Emirates with a population over 2 billion people and a GDP per person.Al-Khaliel airport is an important gateway into Europe to the north east and west of Europe, Asia, and Australia”*
  - Top-words: *people, man, country, city, party, killed, family, agree, wrong, general, children, sex, president, police, working, military, dead, missing, woman, days*
- **Sport**
  - Sample: *“I think it's pretty safe to assume that the guy has a lot of experience with the game. He played in the NHL, and he was a top 10 player on the team for most games last season (he had a goal in his final three playoff series), and he won a Stanley Cup as a rookie this past season (he finished third in the league in points scored, which was good for second in the league) and is still one of the best players in hockey at this stage in the year”*
  - Top-words: *time, game, back, season, play, didn, team, guy, field, night, games, left, 12, title, won, saturday, playing, great, day, wasn*

# Qualitative results: ag_news

- **Business**
  - Sample: *“I am still in shock after hearing of that.It's a pretty big deal. It happened last month. They are trying to get the money out of the company by selling their stock for profit so they can sell more shares and buy more shares at higher prices (which I think would have helped with the stock market) and it was reported as an ""investment fraud"" by the SEC which has been going on all over this subreddit for months, but no one ever seems to care much”*
  - Top-words: *money, buy, care, doesn, things, deal, pay, worth, business, car, biggest, interested, month, trade, don, compagny, happened, store, kind, price*
- **Sci/Tech**
  - Sample: *“2K Games' Dark Souls 3 is coming to PC, Mac \& Linux in the near future.The new game will launch for free on PC, Mac \& Linux and Xbox One, PlayStation 5 and Microsoft Windows, as well. It'll come out sometime during this week, with an official release expected soon thereafter, though we don't yet know what it will be called or where exactly you're getting the title. We also have some news from Sony that's not quite so surprising”*
  - Top-words: *ve, ll, idea, phone, internet, make, system, video, online, life, understand, version, pc, found, 13, thing, computer, lot, hard, issue, people, work, information, future*

# Three Bricks to Consolidate Watermarks for Large Language Models

# Language Model Watermarking



Watermark window

Create seed s = hash($t_{-1}$ ,.., $t_{-h}$)
from h previous tokens

The past 3 years of work in NLP have been ?

$t_{-3}$ $t_{-2}$ $t_{-1}$

Context

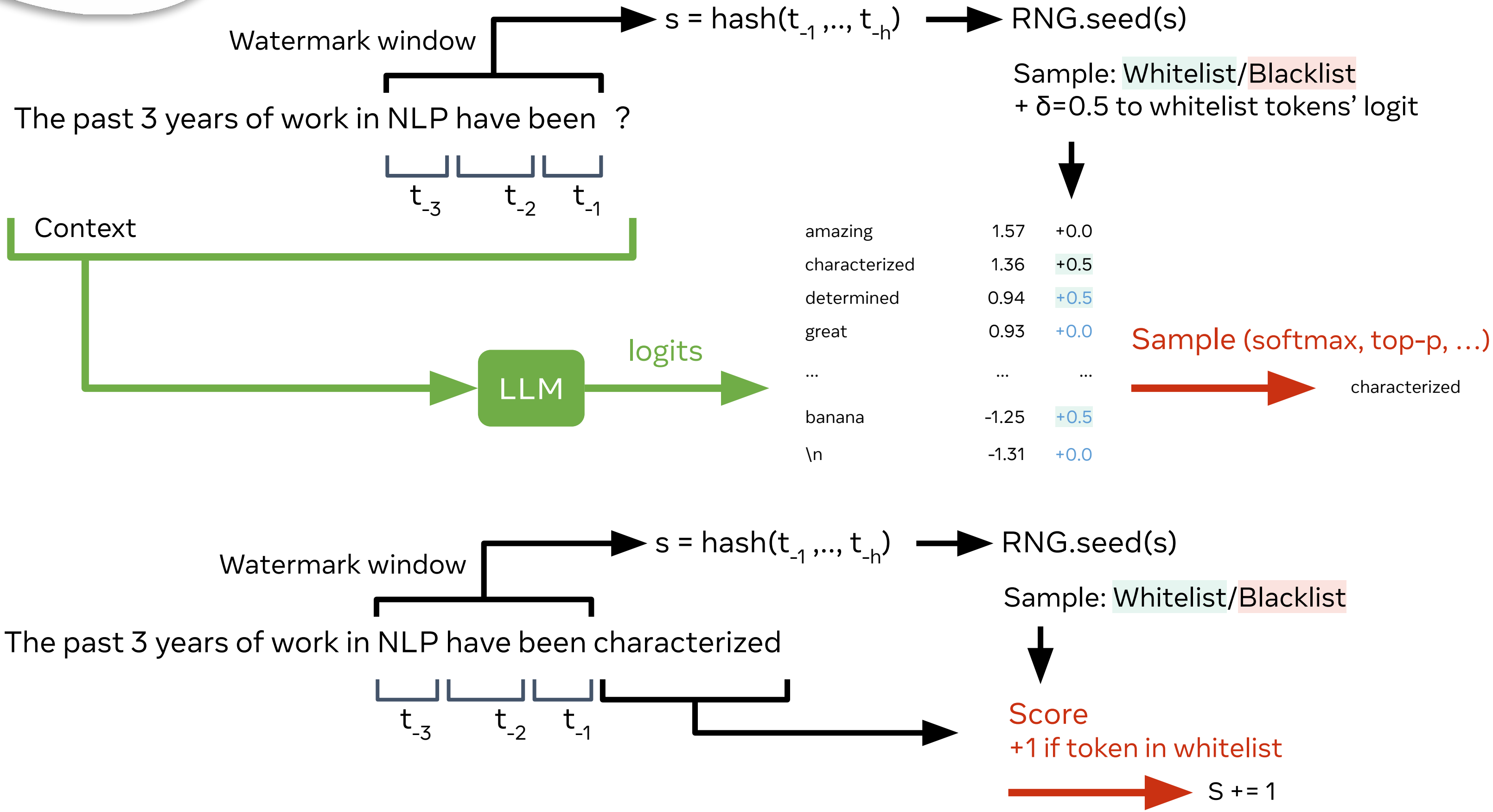
LLM

logits

| | |
|---|---|
| amazing | 1.57 |
| characterized | 1.36 |
| determined | 0.94 |
| great | 0.93 |
| ... | ... |
| banana | -1.25 |
| \n | -1.31 |

Use *s* to seed RNG
Use RNG to sample differently

# A Watermark for Large Language Models



Watermark window

$s = \mathrm{hash}(t_{-1}, .., t_{-h})$ → RNG.seed(s)

Sample: Whitelist/Blacklist
+ δ=0.5 to whitelist tokens' logit

The past 3 years of work in NLP have been ?

$t_{-3}$ $t_{-2}$ $t_{-1}$

Context

LLM

logits

| | | |
|---|---|---|
| amazing | 1.57 | +0.0 |
| characterized | 1.36 | +0.5 |
| determined | 0.94 | +0.5 |
| great | 0.93 | +0.0 |
| ... | ... | ... |
| banana | -1.25 | +0.5 |
| \n | -1.31 | +0.0 |

Sample (softmax, top-p, ...)

characterized

Watermark window

$s = \mathrm{hash}(t_{-1}, .., t_{-h})$ → RNG.seed(s)

Sample: Whitelist/Blacklist

The past 3 years of work in NLP have been characterized

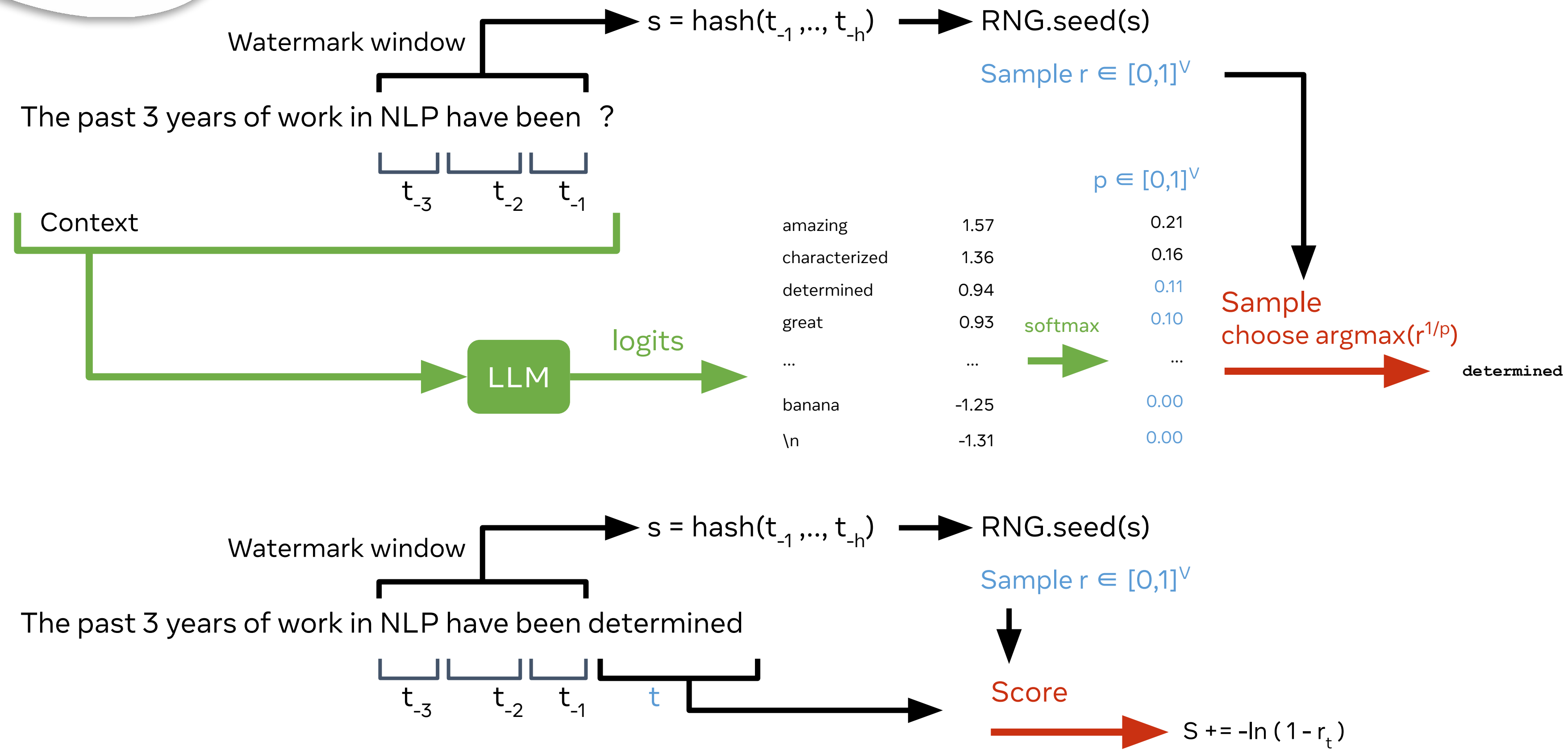
$t_{-3}$ $t_{-2}$ $t_{-1}$

Score
+1 if token in whitelist

S += 1

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, Tom Goldstein. "A Watermark for Large Language Models". 2023

# Watermarking GPT Outputs



Watermark window

s = hash($t_{-1}$ ,.., $t_{-h}$)

RNG.seed(s)

Sample $r \in [0,1]^V$

The past 3 years of work in NLP have been ?

$t_{-3}$ $t_{-2}$ $t_{-1}$

Context

LLM

logits

| | | $p \in [0,1]^V$ |
|---|---|---|
| amazing | 1.57 | 0.21 |
| characterized | 1.36 | 0.16 |
| determined | 0.94 | 0.11 |
| great | 0.93 | 0.10 |
| ... | ... | ... |
| banana | -1.25 | 0.00 |
| \n | -1.31 | 0.00 |

softmax

Sample

choose argmax($r^{1/p}$)

`determined`

Watermark window

s = hash($t_{-1}$ ,.., $t_{-h}$)

RNG.seed(s)

Sample $r \in [0,1]^V$

The past 3 years of work in NLP have been determined

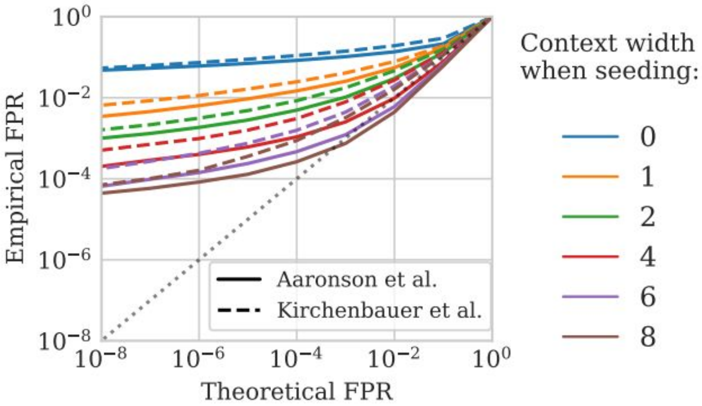
$t_{-3}$ $t_{-2}$ $t_{-1}$ $t$

Score

$S \mathrel{+}= -\ln(1 - r_t)$

# Watermark detection

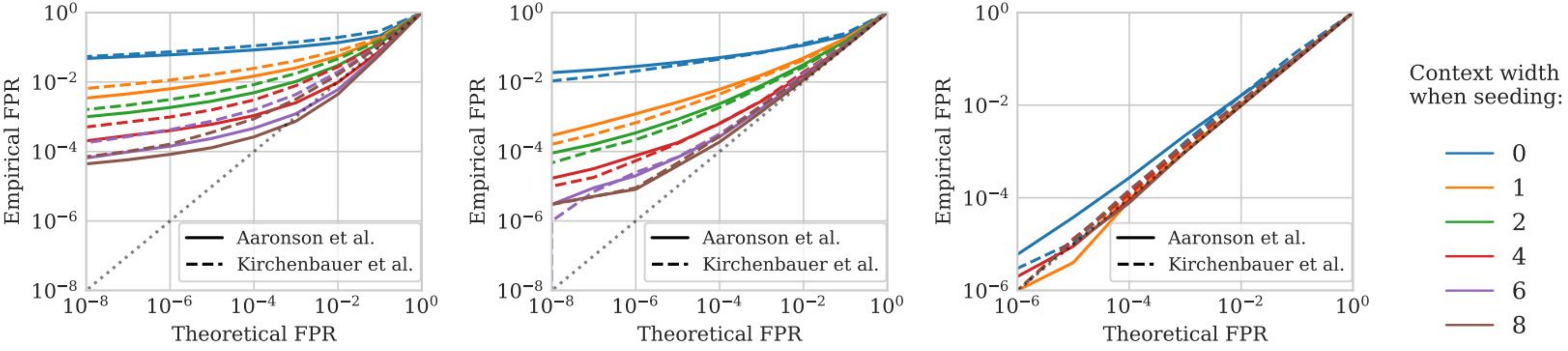
- Test the hypothesis that the text is natural using a statistical test
  - Does the sample differ **significantly** from the standard distribution
- Compute p-value, probability of observing this deviation randomly
  - Rely on **Gaussian assumption**
  - Set a **false positive rate (FPR)** and set threshold accordingly



(a) $Z$-tests

# Correcting FPR

1. Introduce non-asymptotical statistical tests
2. Ignore already seen context windows that share the same secret (=> pseudo-random variable)
   - ○ Because some subsequences are frequent (bulleted list, etc)



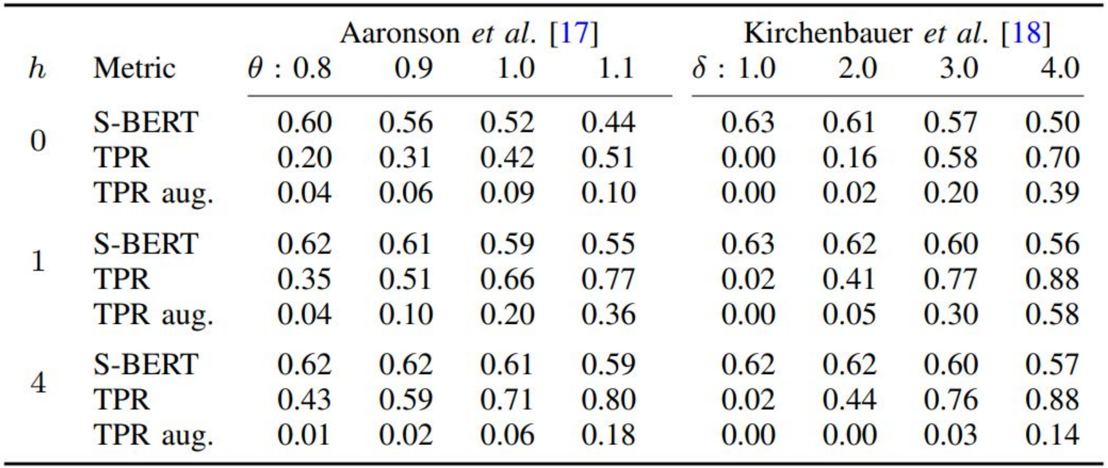
(a) $Z$-tests

(b) Tests of III-B

(c) Tests of III-B, rectified with III-C

# Robustness analysis

- Evaluating detection (TPR) for both methods using different strength/window size
- Kirchenbauer has finer control over trade-off
- Aaronson is more consistent but fails to achieve very high TPR
- Small window size yields biased and repetitive texts but more robust detection (re-synchronization)



| | | Aaronson *et al.* [17] | | | | Kirchenbauer *et al.* [18] | | | |
|---|---|---|---|---|---|---|---|---|---|
| $h$ | Metric | $\theta$ : 0.8 | 0.9 | 1.0 | 1.1 | $\delta$ : 1.0 | 2.0 | 3.0 | 4.0 |
| 0 | S-BERT | 0.60 | 0.56 | 0.52 | 0.44 | 0.63 | 0.61 | 0.57 | 0.50 |
| | TPR | 0.20 | 0.31 | 0.42 | 0.51 | 0.00 | 0.16 | 0.58 | 0.70 |
| | TPR aug. | 0.04 | 0.06 | 0.09 | 0.10 | 0.00 | 0.02 | 0.20 | 0.39 |
| 1 | S-BERT | 0.62 | 0.61 | 0.59 | 0.55 | 0.63 | 0.62 | 0.60 | 0.56 |
| | TPR | 0.35 | 0.51 | 0.66 | 0.77 | 0.02 | 0.41 | 0.77 | 0.88 |
| | TPR aug. | 0.04 | 0.10 | 0.20 | 0.36 | 0.00 | 0.05 | 0.30 | 0.58 |
| 4 | S-BERT | 0.62 | 0.62 | 0.61 | 0.59 | 0.62 | 0.62 | 0.60 | 0.57 |
| | TPR | 0.43 | 0.59 | 0.71 | 0.80 | 0.02 | 0.44 | 0.76 | 0.88 |
| | TPR aug. | 0.01 | 0.02 | 0.06 | 0.18 | 0.00 | 0.00 | 0.03 | 0.14 |

# Free-form evaluation

- Evaluate true usefulness of watermarked output using NLP tasks
- Compare samples to a set of references (because WM affects sampling, not likelihood)
- WM does not degrade results much
  - Sampling-based generation

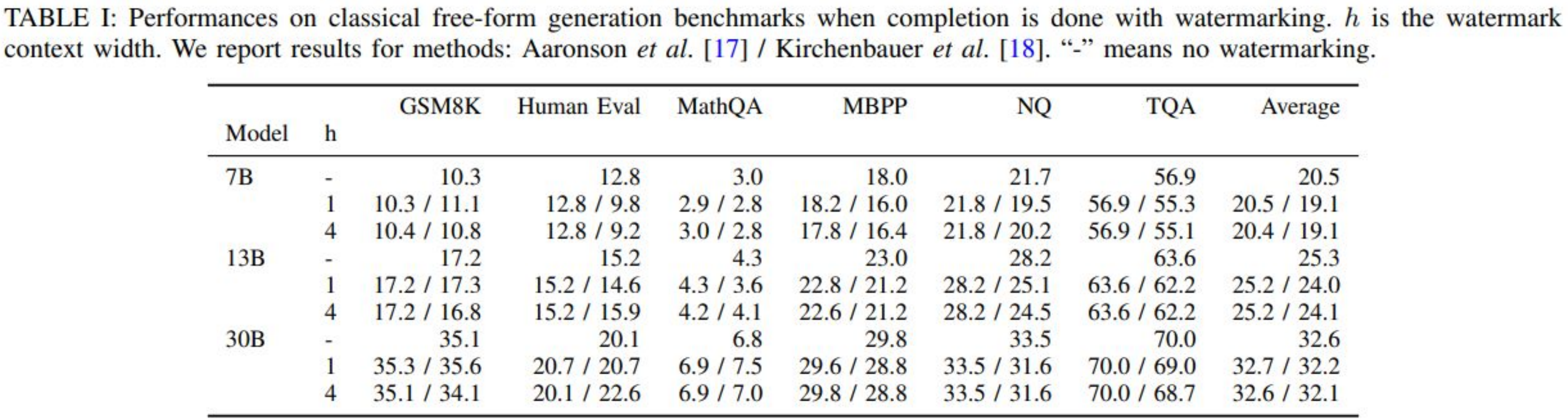


TABLE I: Performances on classical free-form generation benchmarks when completion is done with watermarking. *h* is the watermark context width. We report results for methods: Aaronson *et al.* [17] / Kirchenbauer *et al.* [18]. "-" means no watermarking.

| | | GSM8K | Human Eval | MathQA | MBPP | NQ | TQA | Average |
|---|---|---|---|---|---|---|---|---|
| Model | h | | | | | | | |
| 7B | - | 10.3 | 12.8 | 3.0 | 18.0 | 21.7 | 56.9 | 20.5 |
| | 1 | 10.3 / 11.1 | 12.8 / 9.8 | 2.9 / 2.8 | 18.2 / 16.0 | 21.8 / 19.5 | 56.9 / 55.3 | 20.5 / 19.1 |
| | 4 | 10.4 / 10.8 | 12.8 / 9.2 | 3.0 / 2.8 | 17.8 / 16.4 | 21.8 / 20.2 | 56.9 / 55.1 | 20.4 / 19.1 |
| 13B | - | 17.2 | 15.2 | 4.3 | 23.0 | 28.2 | 63.6 | 25.3 |
| | 1 | 17.2 / 17.3 | 15.2 / 14.6 | 4.3 / 3.6 | 22.8 / 21.2 | 28.2 / 25.1 | 63.6 / 62.2 | 25.2 / 24.0 |
| | 4 | 17.2 / 16.8 | 15.2 / 15.9 | 4.2 / 4.1 | 22.6 / 21.2 | 28.2 / 24.5 | 63.6 / 62.2 | 25.2 / 24.1 |
| 30B | - | 35.1 | 20.1 | 6.8 | 29.8 | 33.5 | 70.0 | 32.6 |
| | 1 | 35.3 / 35.6 | 20.7 / 20.7 | 6.9 / 7.5 | 29.6 / 28.8 | 33.5 / 31.6 | 70.0 / 69.0 | 32.7 / 32.2 |
| | 4 | 35.1 / 34.1 | 20.1 / 22.6 | 6.9 / 7.0 | 29.8 / 28.8 | 33.5 / 31.6 | 70.0 / 68.7 | 32.6 / 32.1 |

# Multi-bit watermarking

- Zero-bit watermarking: is the text marked?
- Multi-bit watermarking: if marked, which key (user)?
  - Add a binary message (version of the LM, authorship, …)
- Requires to compute M detections
  - Selecting key as a circular shift of m indices allows to run parallel detection
- Sufficient performance to be dissuasive

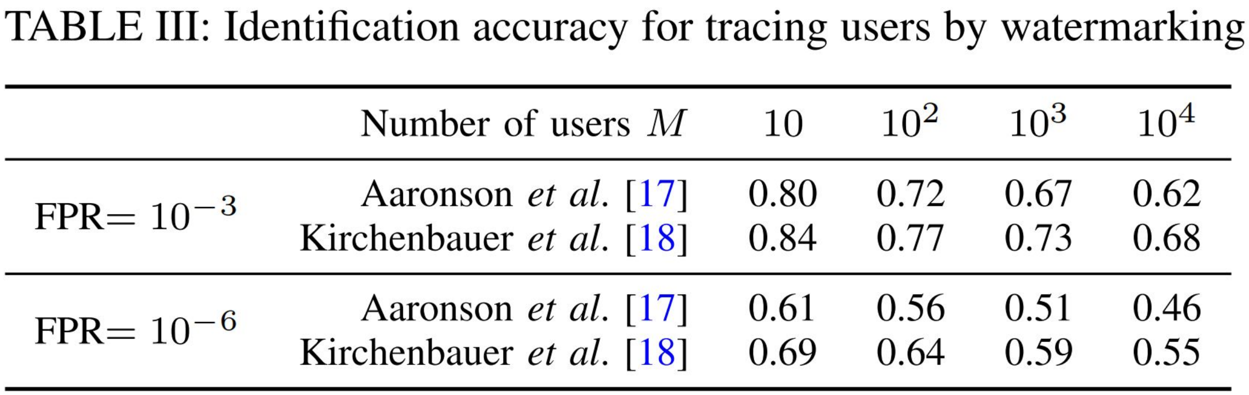
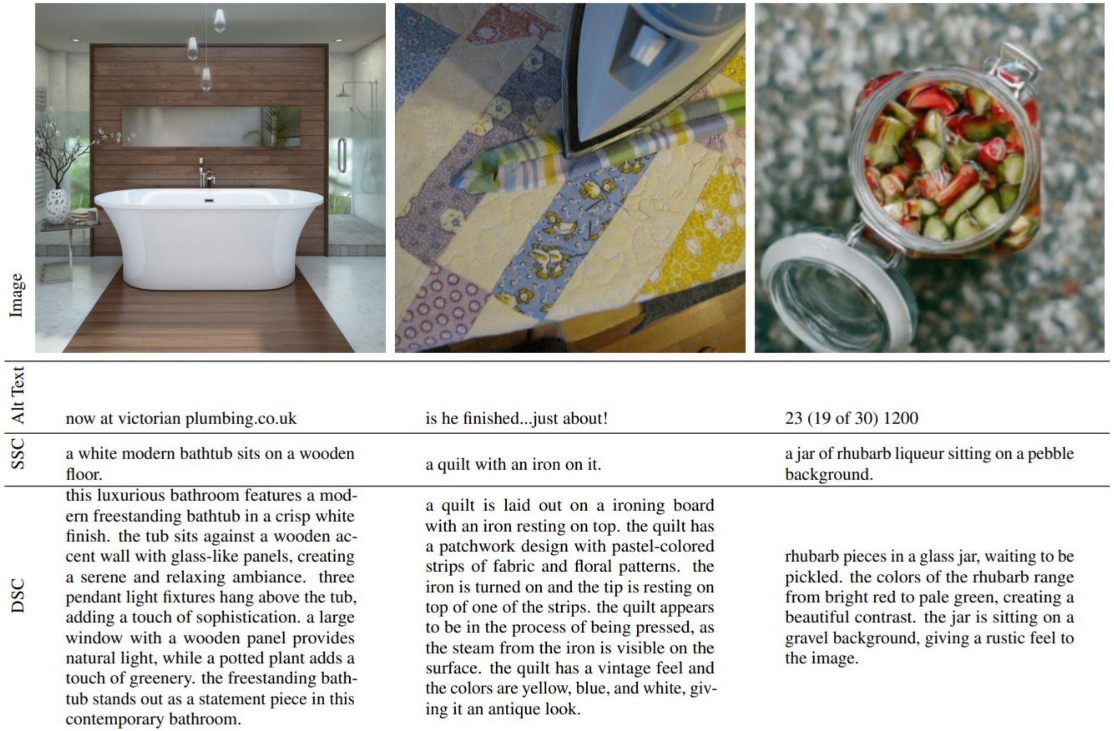


TABLE III: Identification accuracy for tracing users by watermarking

| | Number of users $M$ | 10 | $10^2$ | $10^3$ | $10^4$ |
|---|---|---|---|---|---|
| FPR$= 10^{-3}$ | Aaronson *et al.* [17] | 0.80 | 0.72 | 0.67 | 0.62 |
| | Kirchenbauer *et al.* [18] | 0.84 | 0.77 | 0.73 | 0.68 |
| FPR$= 10^{-6}$ | Aaronson *et al.* [17] | 0.61 | 0.56 | 0.51 | 0.46 |
| | Kirchenbauer *et al.* [18] | 0.69 | 0.64 | 0.59 | 0.55 |

# DALL-E 3



| | | | |
|---|---|---|---|
| Image | | | |
| Alt Text | now at victorian plumbing.co.uk | is he finished...just about! | 23 (19 of 30) 1200 |
| SSC | a white modern bathtub sits on a wooden floor. | a quilt with an iron on it. | a jar of rhubarb liqueur sitting on a pebble background. |
| DSC | this luxurious bathroom features a modern freestanding bathtub in a crisp white finish. the tub sits against a wooden accent wall with glass-like panels, creating a serene and relaxing ambiance. three pendant light fixtures hang above the tub, adding a touch of sophistication. a large window with a wooden panel provides natural light, while a potted plant adds a touch of greenery. the freestanding bathtub stands out as a statement piece in this contemporary bathroom. | a quilt is laid out on a ironing board with an iron resting on top. the quilt has a patchwork design with pastel-colored strips of fabric and floral patterns. the iron is turned on and the tip is resting on top of one of the strips. the quilt appears to be in the process of being pressed, as the steam from the iron is visible on the surface. the quilt has a vintage feel and the colors are yellow, blue, and white, giving it an antique look. | rhubarb pieces in a glass jar, waiting to be pickled. the colors of the rhubarb range from bright red to pale green, creating a beautiful contrast. the jar is sitting on a gravel background, giving a rustic feel to the image. |

**Figure 3** – Examples of alt-text accompanying selected images scraped from the internet, short synthetic captions (SSC), and descriptive synthetic captions (DSC).

# DALL-E 3

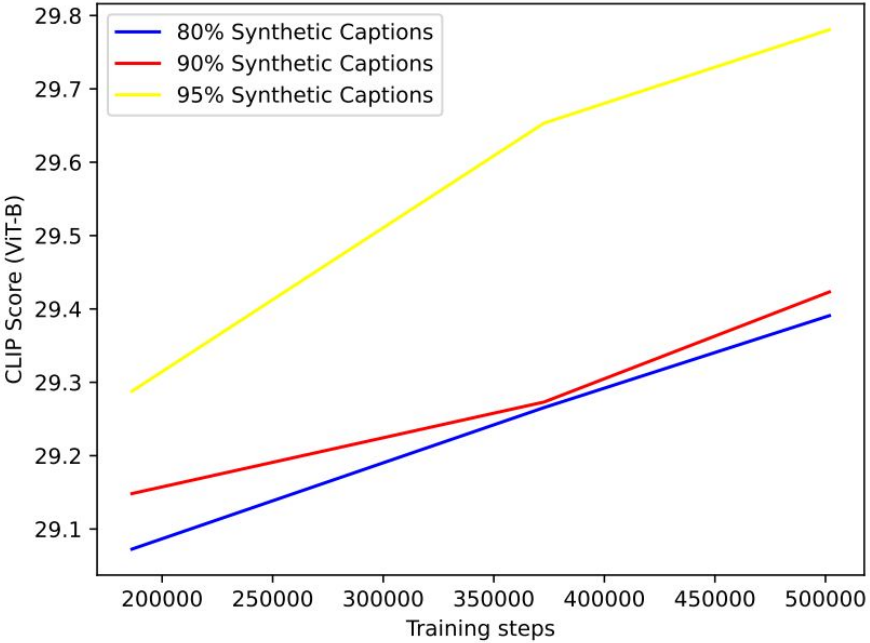
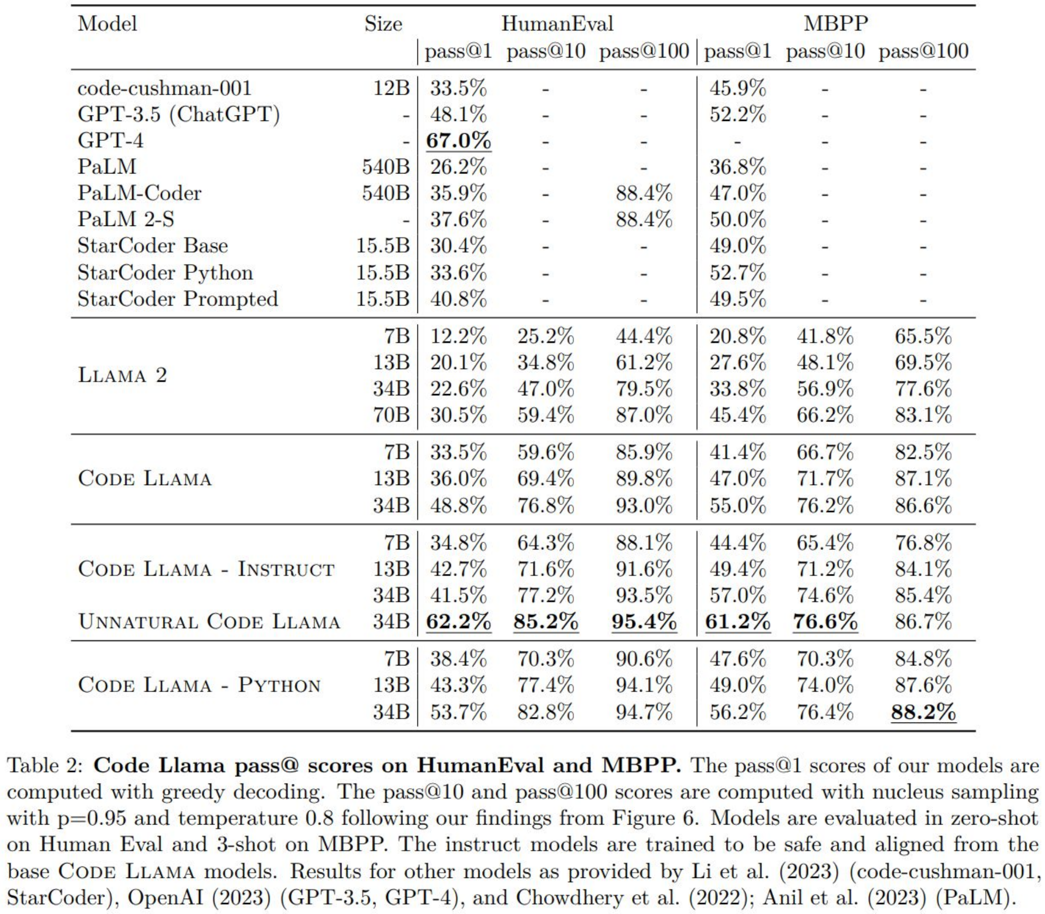


**Figure 5** – CLIP scores for text-to-image models trained on various blending ratios of descriptive synthetic captions and ground-truth captions. Evaluation performed using ground truth captions.

# Unnatural Code Llama



| Model | Size | HumanEval | | | MBPP | | |
|---|---|---|---|---|---|---|---|
| | | pass@1 | pass@10 | pass@100 | pass@1 | pass@10 | pass@100 |
| code-cushman-001 | 12B | 33.5% | - | - | 45.9% | - | - |
| GPT-3.5 (ChatGPT) | - | 48.1% | - | - | 52.2% | - | - |
| GPT-4 | - | **67.0%** | - | - | - | - | - |
| PaLM | 540B | 26.2% | - | - | 36.8% | - | - |
| PaLM-Coder | 540B | 35.9% | - | 88.4% | 47.0% | - | - |
| PaLM 2-S | - | 37.6% | - | 88.4% | 50.0% | - | - |
| StarCoder Base | 15.5B | 30.4% | - | - | 49.0% | - | - |
| StarCoder Python | 15.5B | 33.6% | - | - | 52.7% | - | - |
| StarCoder Prompted | 15.5B | 40.8% | - | - | 49.5% | - | - |
| Llama 2 | 7B | 12.2% | 25.2% | 44.4% | 20.8% | 41.8% | 65.5% |
| | 13B | 20.1% | 34.8% | 61.2% | 27.6% | 48.1% | 69.5% |
| | 34B | 22.6% | 47.0% | 79.5% | 33.8% | 56.9% | 77.6% |
| | 70B | 30.5% | 59.4% | 87.0% | 45.4% | 66.2% | 83.1% |
| Code Llama | 7B | 33.5% | 59.6% | 85.9% | 41.4% | 66.7% | 82.5% |
| | 13B | 36.0% | 69.4% | 89.8% | 47.0% | 71.7% | 87.1% |
| | 34B | 48.8% | 76.8% | 93.0% | 55.0% | 76.2% | 86.6% |
| Code Llama - Instruct | 7B | 34.8% | 64.3% | 88.1% | 44.4% | 65.4% | 76.8% |
| | 13B | 42.7% | 71.6% | 91.6% | 49.4% | 71.2% | 84.1% |
| | 34B | 41.5% | 77.2% | 93.5% | 57.0% | 74.6% | 85.4% |
| Unnatural Code Llama | 34B | **62.2%** | **85.2%** | **95.4%** | **61.2%** | **76.6%** | 86.7% |
| Code Llama - Python | 7B | 38.4% | 70.3% | 90.6% | 47.6% | 70.3% | 84.8% |
| | 13B | 43.3% | 77.4% | 94.1% | 49.0% | 74.0% | 87.6% |
| | 34B | 53.7% | 82.8% | 94.7% | 56.2% | 76.4% | 88.2% |

Table 2: **Code Llama pass@ scores on HumanEval and MBPP.** The pass@1 scores of our models are computed with greedy decoding. The pass@10 and pass@100 scores are computed with nucleus sampling with p=0.95 and temperature 0.8 following our findings from Figure 6. Models are evaluated in zero-shot on Human Eval and 3-shot on MBPP. The instruct models are trained to be safe and aligned from the base Code Llama models. Results for other models as provided by Li et al. (2023) (code-cushman-001, StarCoder), OpenAI (2023) (GPT-3.5, GPT-4), and Chowdhery et al. (2022); Anil et al. (2023) (PaLM).