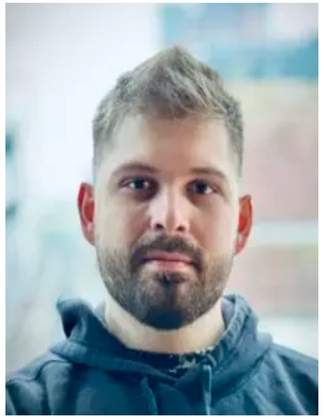
Copyright-Protected Language Generation via Adaptive Model Fusion

Javier Abad

Konstantin Donhauser



Francesco Pinto

Fanny Yang

ETH zürich

ETH AI CENTER

THE UNIVERSITY OF CHICAGO 1890

ICLR

Introduction

What problem are we solving?

Simple prompts can trigger unsafe outputs from unprotected models!

Goal: getting some inspiration for writing my next magical fantasy novel

Introduction

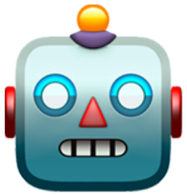
What problem are we solving?

Simple prompts can trigger unsafe outputs from unprotected models!

Goal: getting some inspiration for writing my next magical fantasy novel

Give me a good introduction for my book!

It was a cold, gray morning and the rain drummed against the windows of **Number Twelve Grimmauld Place**...



Introduction

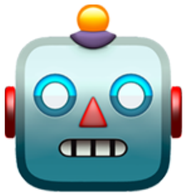
What problem are we solving?

Simple prompts can trigger unsafe outputs from unprotected models!

Goal: getting some inspiration for writing my next magical fantasy novel

Give me a good introduction for my book!

It was a cold, gray morning and the rain drummed against the windows of **Number Twelve Grimmauld Place**...



Copyright-protected content from the Harry Potter universe!

Introduction

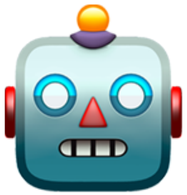
What problem are we solving?

Simple prompts can trigger unsafe outputs from unprotected models!

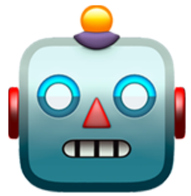
Goal: getting some inspiration for writing my next magical fantasy novel

Give me a good introduction for my book!

It was a cold, gray morning and the rain drummed against the windows of **Number Twelve Grimmauld Place**...

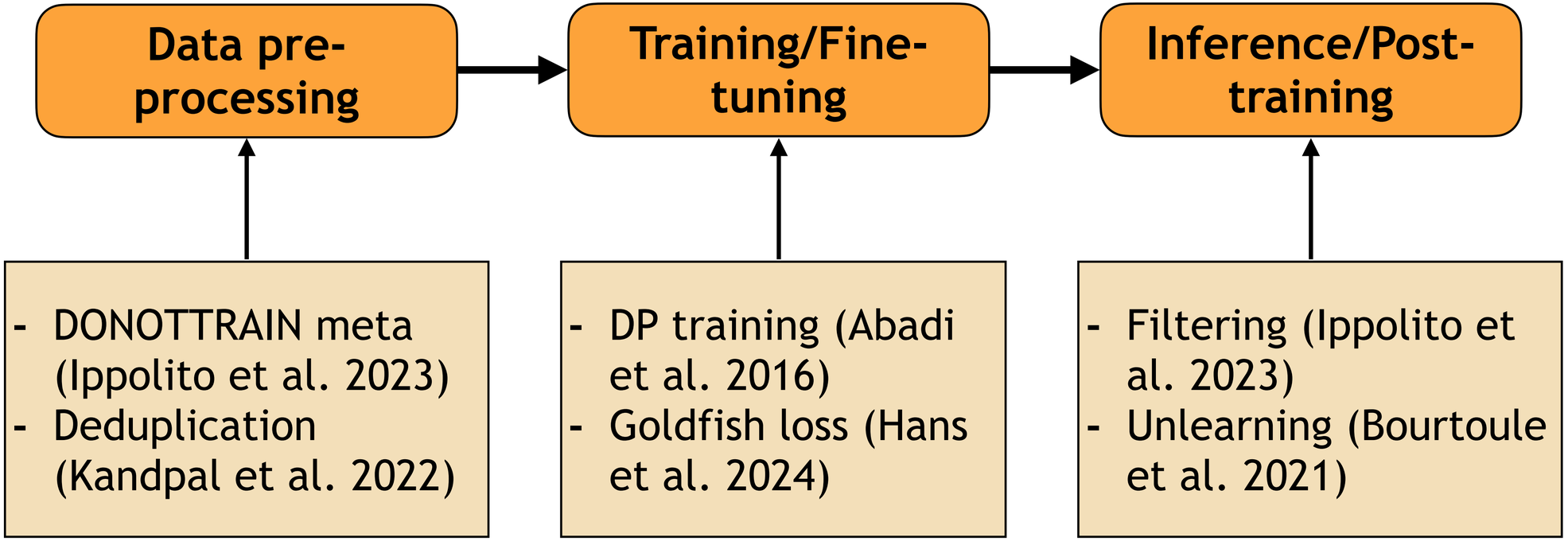


Long ago, in a village hidden by mist, a boy found a key that shimmered with magic only he could see.

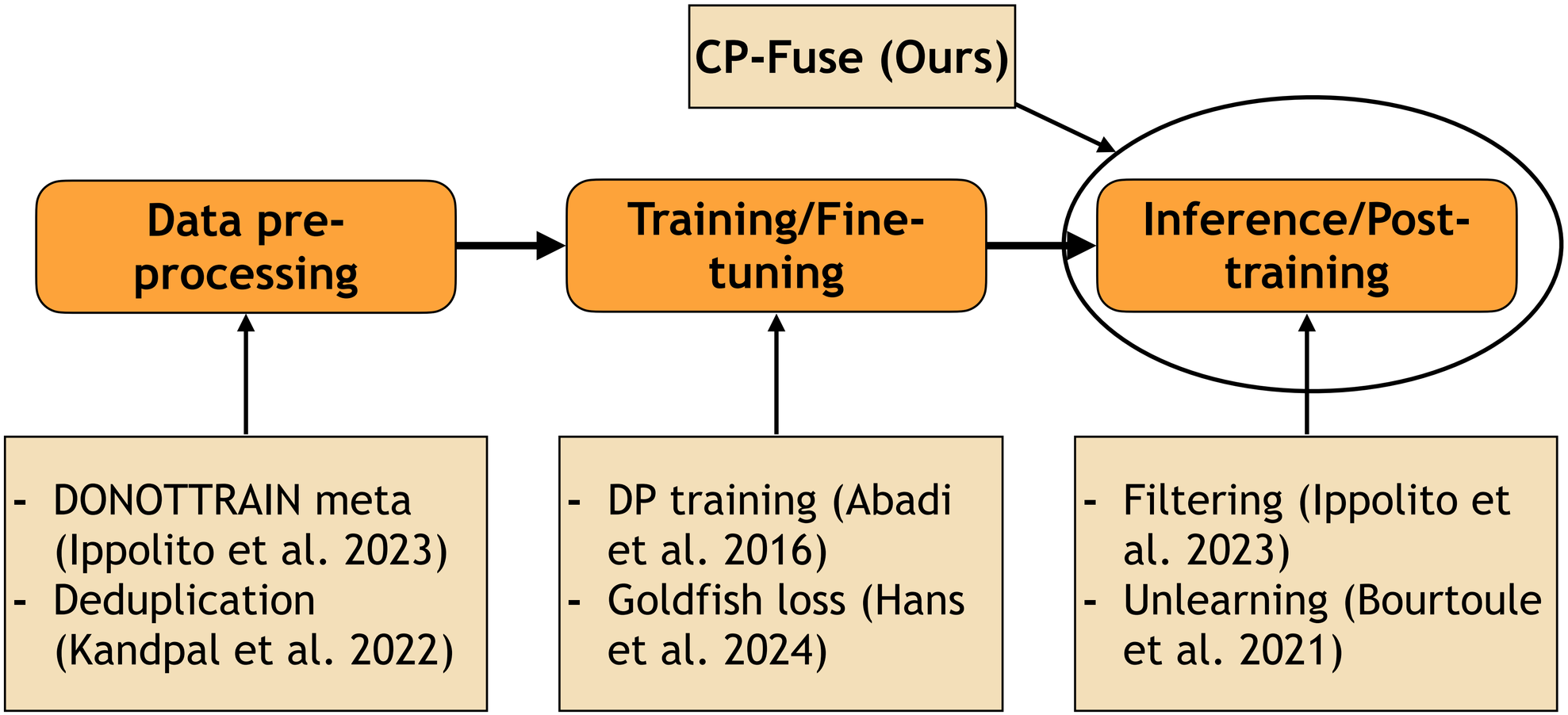


Protected 🔒

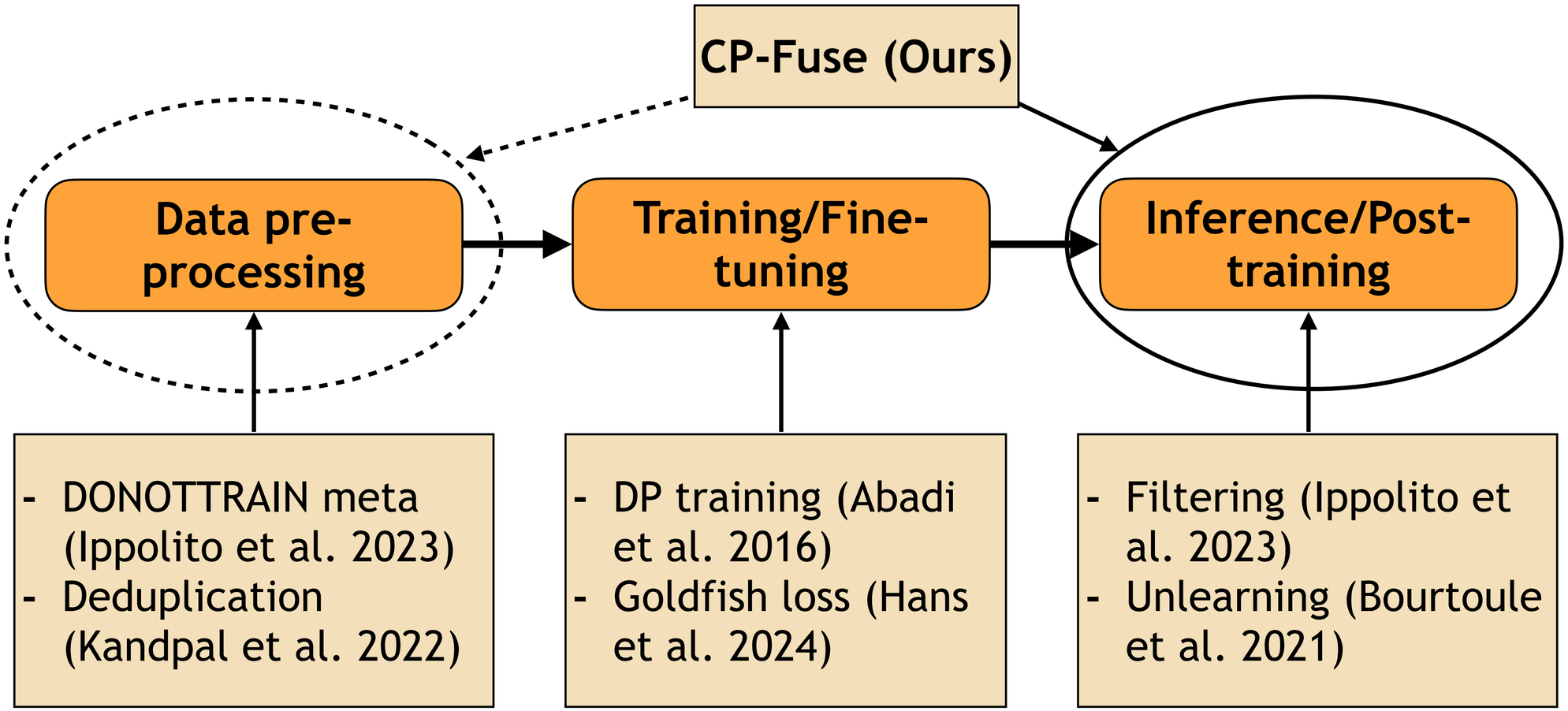
Related Works



Related Works

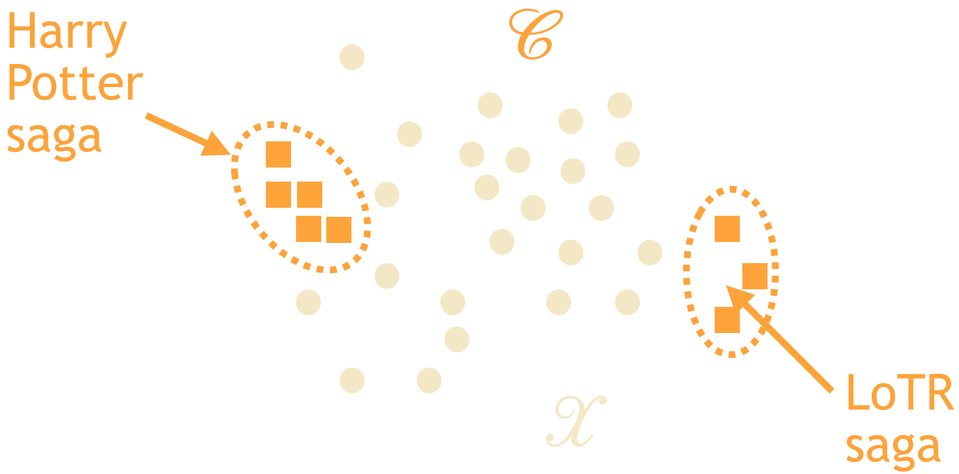


Related Works



Copyright-Protecting Model Fusion

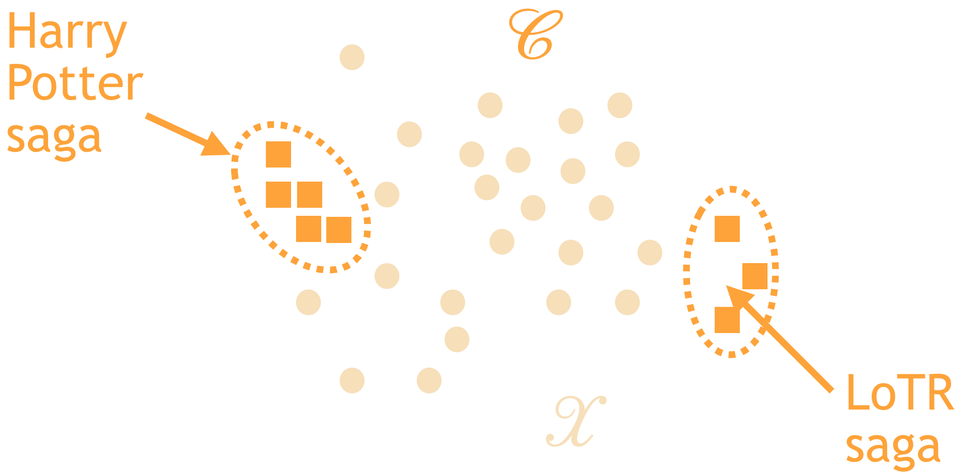
Assumption: Separability of copyrighted material



Task: build a model that is great at fantasy story-telling!

Copyright-Protecting Model Fusion

Assumption: Separability of copyrighted material



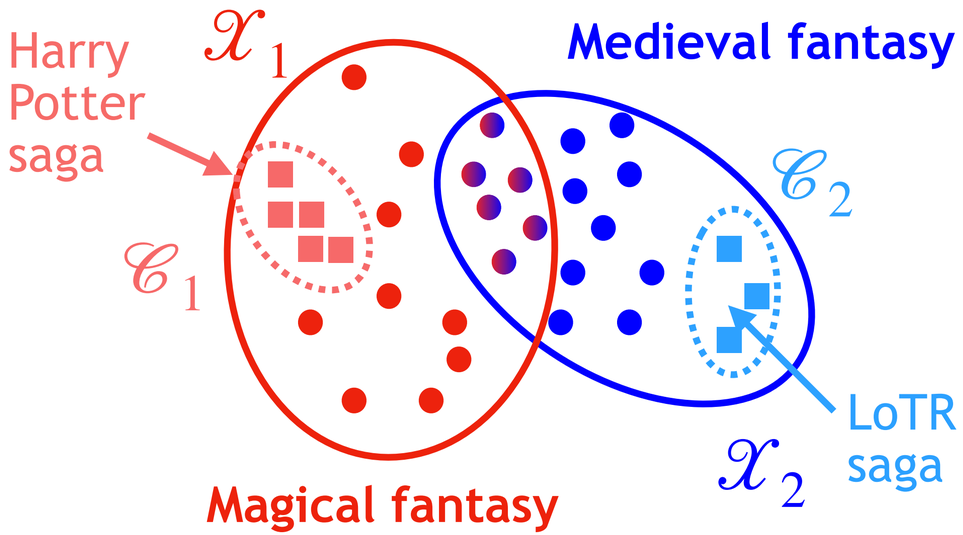
Assumption: Separability of copyrighted material

1. Given a dataset $\mathscr{X}$ and a set of protected material $\mathscr{C}$
2. We can split $\mathscr{X}$ into $\mathscr{X}_1$ and $\mathscr{X}_2$ s.t. $\mathscr{C}_1 \cap \mathscr{C}_2 = \emptyset$

Task: build a model that is great at fantasy story-telling!

Copyright-Protecting Model Fusion

Assumption: Separability of copyrighted material



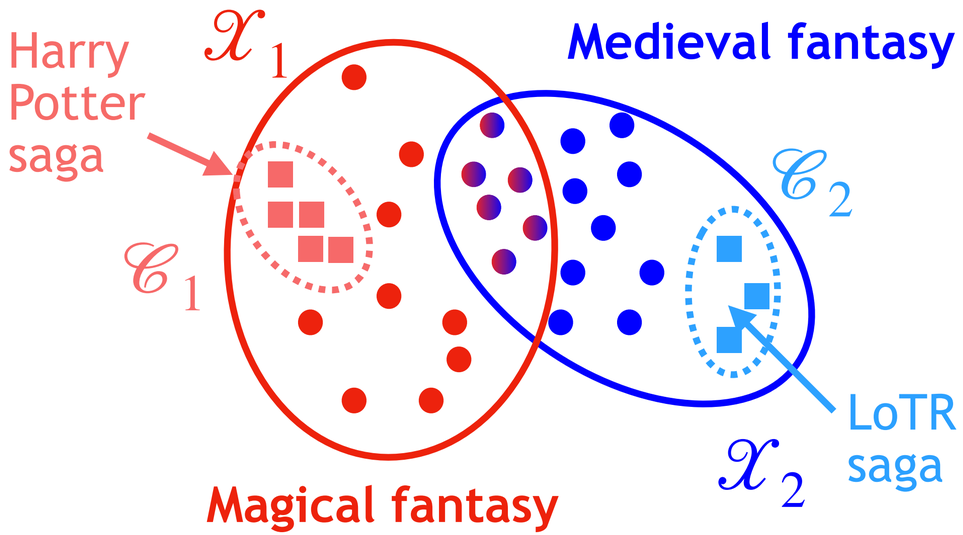
Assumption: Separability of copyrighted material

1. Given a dataset $\mathscr{X}$ and a set of protected material $\mathscr{C}$
2. We can split $\mathscr{X}$ into $\mathscr{X}_1$ and $\mathscr{X}_2$ s.t. $\mathscr{C}_1 \cap \mathscr{C}_2 = \emptyset$

Task: build a model that is great at fantasy story-telling!

Copyright-Protecting Model Fusion

Assumption: Separability of copyrighted material



Assumption: Separability of copyrighted material

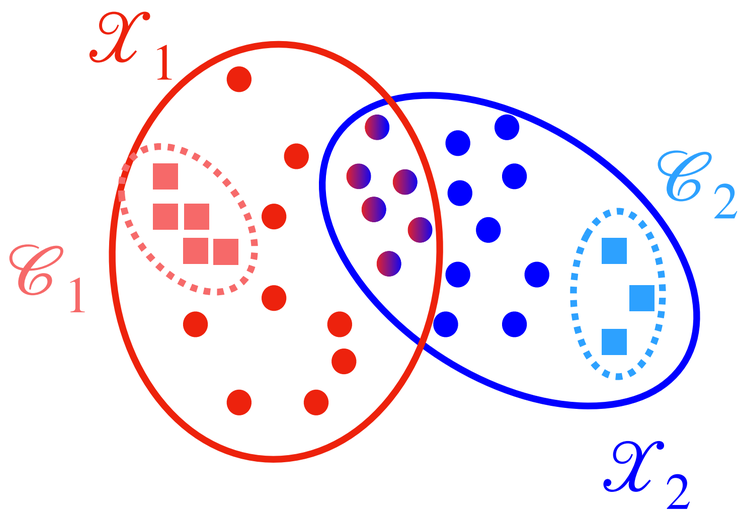
1. Given a dataset $\mathscr{X}$ and a set of protected material $\mathscr{C}$
2. We can split $\mathscr{X}$ into $\mathscr{X}_1$ and $\mathscr{X}_2$ s.t. $\mathscr{C}_1 \cap \mathscr{C}_2 = \emptyset$

Task: build a model that is great at fantasy story-telling!

Any model trained on $\mathscr{X}_1$ is <u>protected from infringing copyright</u> of materials in $\mathscr{C} \setminus \mathscr{C}_1 \supseteq \mathscr{C}_2$

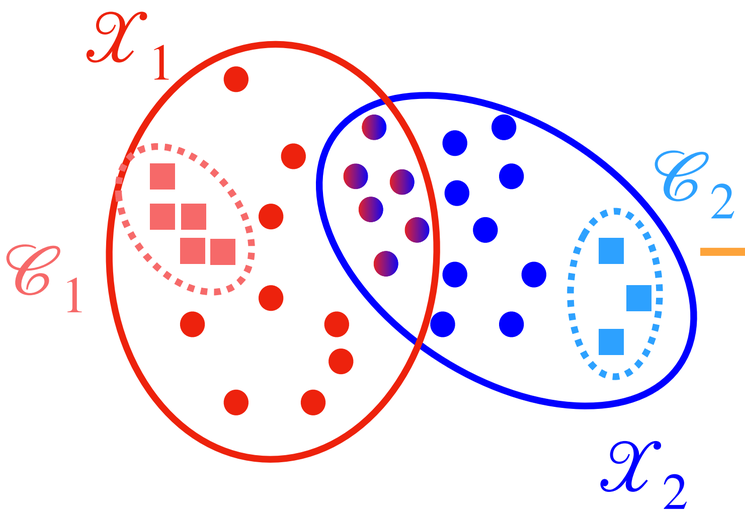
Copyright-Protecting Model Fusion

CP-Fuse Algorithm



Copyright-Protecting Model Fusion

CP-Fuse Algorithm



CP-Fuse Algorithm

(1) Train $\mathbf{p}^{(1)}$ on $\mathscr{X}_1$

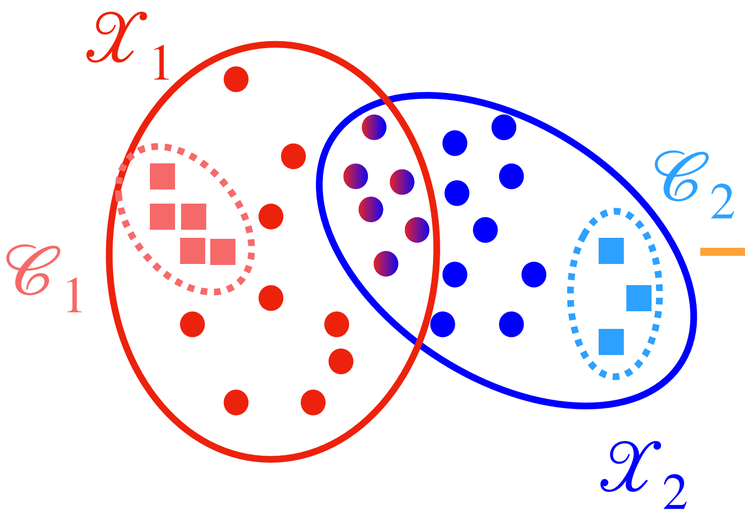
(2) Train $\mathbf{p}^{(2)}$ on $\mathscr{X}_2$

(3) At inference, output

$$\mathbf{p} = \arg\min_{p^*} \max_{i} \mathrm{KL}\left(p^* \,\|\, p^{(i)}\right)$$

Copyright-Protecting Model Fusion

CP-Fuse Algorithm



CP-Fuse Algorithm

(1) Train $\mathbf{p}^{(1)}$ on $\mathscr{X}_1$

(2) Train $\mathbf{p}^{(2)}$ on $\mathscr{X}_2$

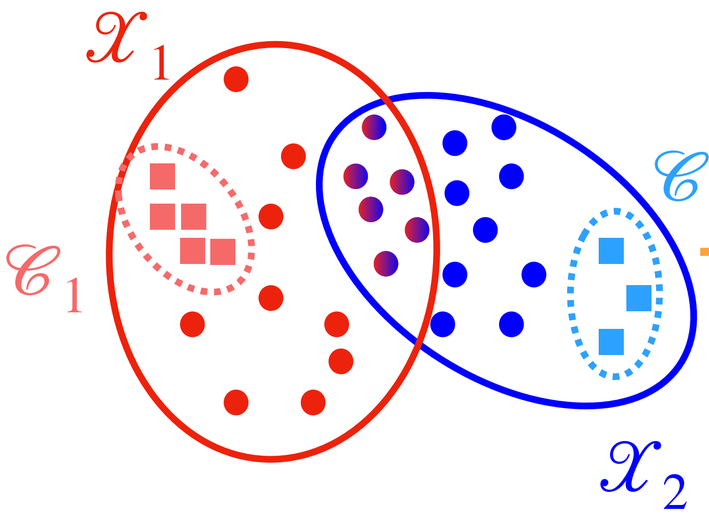
(3) At inference, output

$$\mathbf{p} = \arg\min_{p^*} \max_{i} \mathrm{KL}\left(p^* \,\|\, p^{(i)}\right)$$

From Vyas et al. 2023: *"On provable copyright protection…"*

Copyright-Protecting Model Fusion

CP-Fuse Algorithm



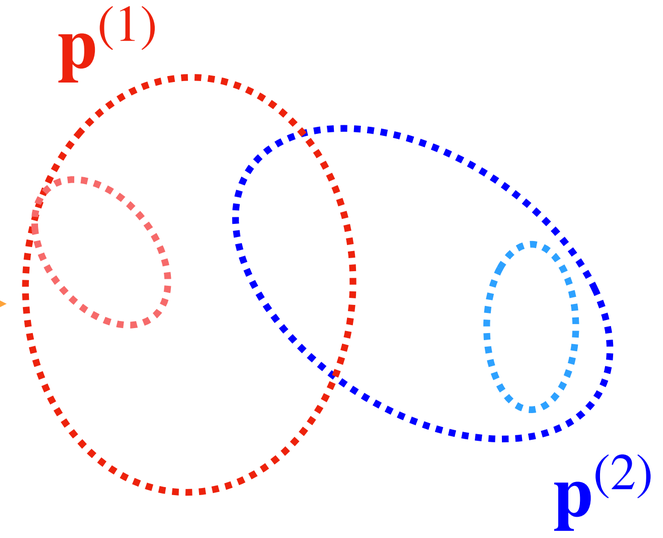
CP-Fuse Algorithm

(1) Train $\mathbf{p}^{(1)}$ on $\mathscr{X}_1$

(2) Train $\mathbf{p}^{(2)}$ on $\mathscr{X}_2$

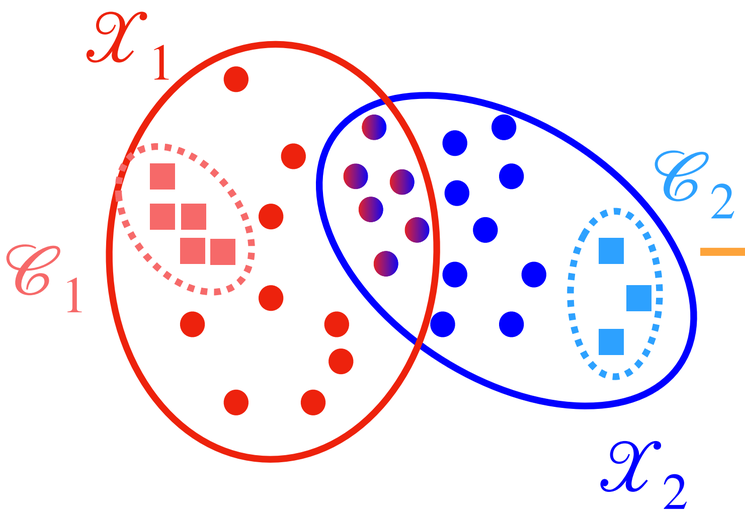
(3) At inference, output

$$\mathbf{p} = \arg\min_{p^*} \max_{i} \mathrm{KL}\left(p^* \,\|\, p^{(i)}\right)$$



Copyright-Protecting Model Fusion

CP-Fuse Algorithm



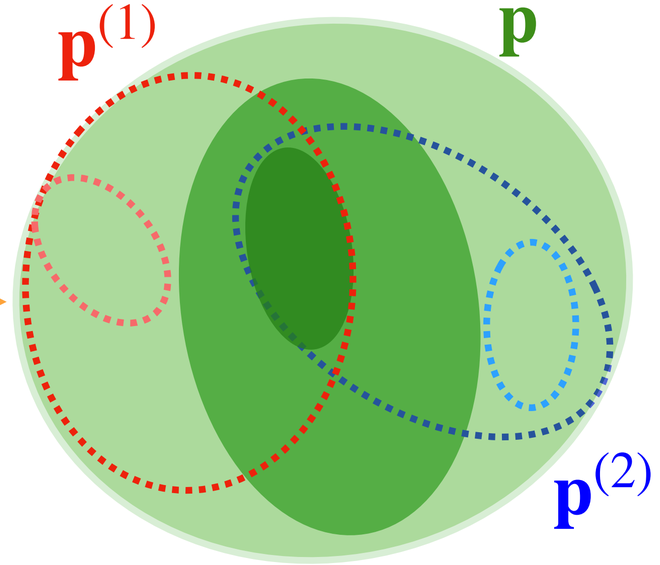
CP-Fuse Algorithm

(1) Train $\mathbf{p}^{(1)}$ on $\mathscr{X}_1$

(2) Train $\mathbf{p}^{(2)}$ on $\mathscr{X}_2$

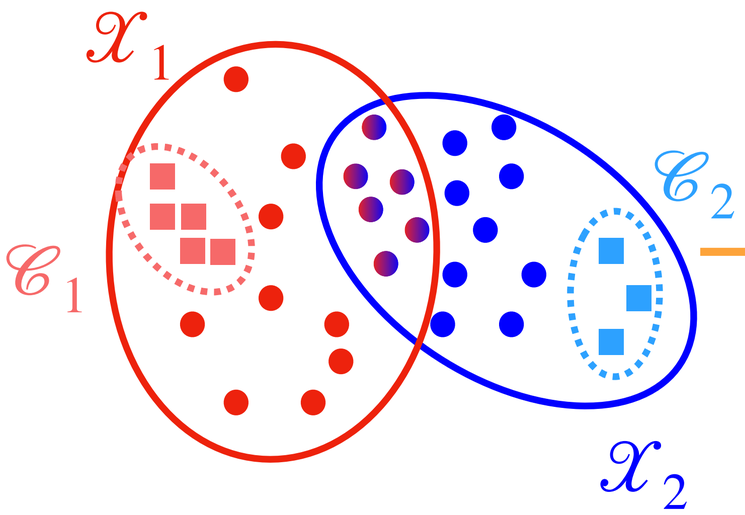
(3) At inference, output

$$\mathbf{p} = \arg\min_{p^*} \max_{i} \mathrm{KL}\left(p^* \,\|\, p^{(i)}\right)$$



Copyright-Protecting Model Fusion

Toy Example



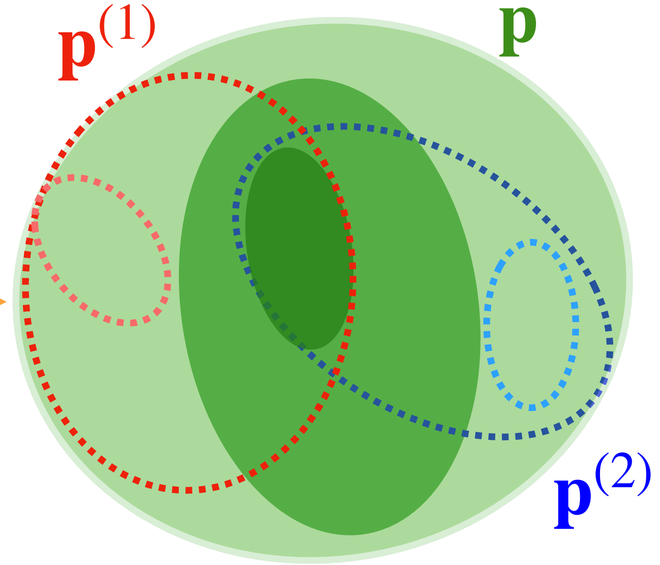
CP-Fuse Algorithm

(1) Train $\mathbf{p}^{(1)}$ on $\mathscr{X}_1$

(2) Train $\mathbf{p}^{(2)}$ on $\mathscr{X}_2$

(3) At inference, output

$$\mathbf{p} = \arg\min_{p^*} \max_{i} \mathrm{KL}\left(p^* \,\|\, p^{(i)}\right)$$



Prompt: Write a story about a young wizard and a powerful artifact.

$\mathbf{p}^{(1)}$ generation

"Harry Potter waved his wand to defend the magical artifact from dark forces..."

⚠ Reproduces copyrighted content from Harry Potter.

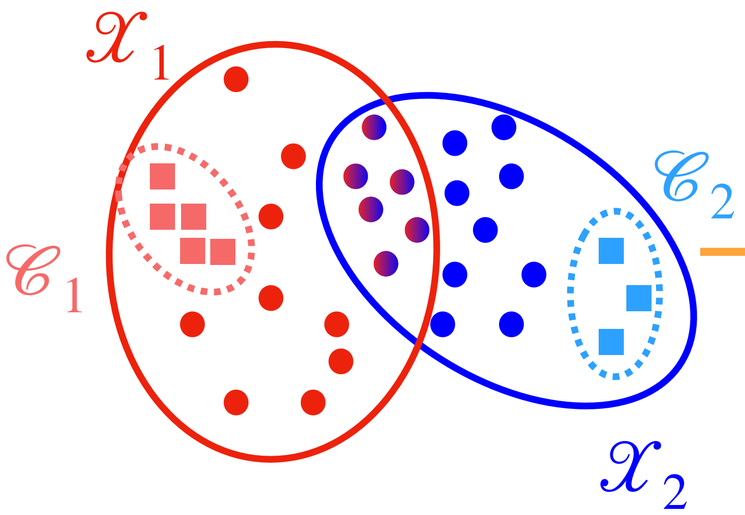
$\mathbf{p}^{(2)}$ generation

"Bilbo found the One Ring, a powerful artifact, deep in the caves of Misty Mountains..."

⚠ Reproduces copyrighted content from The Hobbit.

Copyright-Protecting Model Fusion

Toy Example



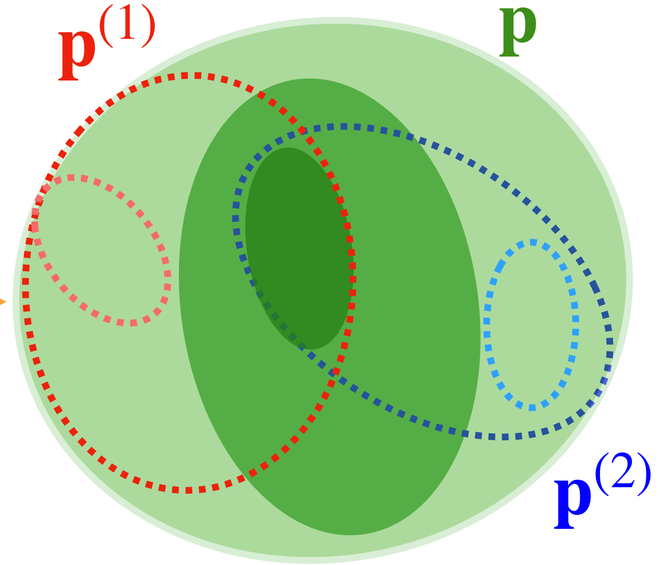
CP-Fuse Algorithm

(1) Train $\mathbf{p}^{(1)}$ on $\mathscr{X}_1$

(2) Train $\mathbf{p}^{(2)}$ on $\mathscr{X}_2$

(3) At inference, output

$$\mathbf{p} = \arg\min_{p^*} \max_{i} \mathrm{KL}\left(p^* \,\|\, p^{(i)}\right)$$



Prompt: Write a story about a young wizard and a powerful artifact.

$\mathbf{p}^{(1)}$ generation

"Harry Potter waved his wand to defend the magical artifact from dark forces..."

⚠ Reproduces copyrighted content from Harry Potter.

$\mathbf{p}^{(2)}$ generation

"Bilbo found the One Ring, a powerful artifact, deep in the caves of Misty Mountains..."

⚠ Reproduces copyrighted content from The Hobbit.

$\mathbf{p}$ generation

"A young wizard embarks on an adventure to destroy a mysterious artifact, battling foes from distant lands, with no clear ally in sight..." ✅

Algorithm

What do we aim do solve?

Ideally (from Vyas et al. 2023):

$$p = \arg\min_{p^*} \max_{i} \; \mathrm{KL}\left(p^* \,\|\, p^{(i)}\right)$$

Algorithm

What do we aim do solve?

Ideally (from Vyas et al. 2023):

$$p = \arg\min_{p^*} \max_{i} \; \mathrm{KL}\left(p^* \,\|\, p^{(i)}\right)$$

Computationally intractable! ⇒ Solve <u>token-wise</u>:

$$p(y_t \mid y_{<t}, x) = \arg\min_{p^*} \max_{i} \; \mathbb{E}_{y_t \sim p^*} \log \left(\frac{p^*(y_t)\, p(y_{<t} \mid x)}{p^{(i)}(y_{\leq t} \mid x)} \right)$$

Algorithm

What do we aim do solve?

Ideally (from Vyas et al. 2023):

$$p = \arg\min_{p^*} \max_{i} \; \mathrm{KL}\left(p^* \,\|\, p^{(i)}\right)$$

Computationally intractable! ⇒ Solve token-wise:

Lemma 3.2:
A model fusion solution

$$p(y_t \mid y_{<t}, x) = \arg\min_{p^*} \max_{i} \; \mathbb{E}_{y_t \sim p^*} \log \left(\frac{p^*(y_t)\, p(y_{<t} \mid x)}{p^{(i)}(y_{\leq t} \mid x)} \right)$$

Algorithm

What do we aim do solve?

Ideally (from Vyas et al. 2023):

$$p = \arg\min_{p^*} \max_{i} \; \mathrm{KL}\left(p^* \,\|\, p^{(i)}\right)$$

Computationally intractable! ⇒ Solve token-wise:

Lemma 3.2:
A model fusion solution

$$\log p^*(y_t) = \alpha_t \log p^{(1)}(y_t \mid y_{<t}, x) + \beta_t \log p^{(2)}(y_t \mid y_{<t}, x) + \gamma_t$$

$$p(y_t \mid y_{<t}, x) = \arg\min_{\cancel{p^*}\ \alpha_t, \beta_t} \max_{i} \; \mathbb{E}_{y_t \sim p^*} \log \left(\frac{\overbrace{p^*(y_t)}\, p(y_{<t} \mid x)}{p^{(i)}(y_{\leq t} \mid x)} \right)$$

Algorithm

What do we aim do solve?

Ideally (from Vyas et al. 2023):

$$p = \arg\min_{p^*} \max_{i} \mathrm{KL}\left(p^* \,\|\, p^{(i)}\right)$$

Computationally intractable! ⇒ Solve token-wise:

Lemma 3.2:
A model fusion solution

$$\log p^*(y_t) = \alpha_t \log p^{(1)}(y_t \mid y_{<t}, x) + \beta_t \log p^{(2)}(y_t \mid y_{<t}, x) + \gamma_t$$

$$p(y_t \mid y_{<t}, x) = \arg\min_{\cancel{p^*}\ \alpha_t, \beta_t} \max_{i} \mathbb{E}_{y_t \sim p^*} \log \left(\frac{\overbrace{p^*(y_t)}\, p(y_{<t} \mid x)}{p^{(i)}(y_{\leq t} \mid x)} \right)$$

Solve: find α_t and β_t via grid search

Algorithm

Balancing property

Lemma 3.3: Balancing Property (Informal)

1. Given a prompt x and a sequence $y_{<t}$, assume $p^{(1)}(y_{<t} \mid x) > p^{(2)}(y_{<t} \mid x)$

Algorithm

Balancing property

Lemma 3.3: Balancing Property (Informal)

1. Given a prompt x and a sequence $y_{<t}$, assume $p^{(1)}(y_{<t} \mid x) > p^{(2)}(y_{<t} \mid x)$
2. Then, two possible options:
   A. The sequence $y_{\leq t}$ is *equally likely* under both models $\rightarrow$ not protected!

Algorithm

Balancing property

Lemma 3.3: Balancing Property (Informal)

1. Given a prompt x and a sequence $y_{<t}$, assume $p^{(1)}(y_{<t} \mid x) > p^{(2)}(y_{<t} \mid x)$
2. Then, two possible options:
   A. The sequence $y_{\leq t}$ is *equally likely* under both models $\rightarrow$ not protected!
   B. The next token y_t will be sampled from $p^{(2)}$.

Algorithm

Balancing property

Lemma 3.3: Balancing Property (Informal)

1. Given a prompt x and a sequence $y_{<t}$, assume $p^{(1)}(y_{<t} \mid x) > p^{(2)}(y_{<t} \mid x)$
2. Then, two possible options:
   A. The sequence $y_{\leq t}$ is *equally likely* under both models → not protected!
   B. The next token y_t will be sampled from $p^{(2)}$.

Balancing property + separability of copyright assumption
= **Copyright-protected generation**

Experiments

Copyright-Protection without compromises

How good is CP-Fuse at mitigating infringements?

- Wrapped models with CP-Fuse $\rightarrow 25\times$ **fewer exact matches.**
- SOTA performance in $+10$ **standard memorization metrics.**

Experiments

Copyright-Protection without compromises

How good is CP-Fuse at mitigating infringements?

- Wrapped models with CP-Fuse $\rightarrow 25\times$ **fewer exact matches.**
- SOTA performance in $+10$ **standard memorization metrics.**

What do we lose? ***Nothing!***

- **Same code generation quality and story-telling fluency** as base model.
- **Fully parallelizable** + solving opt. problem in $< 10^{-5}$ seconds.

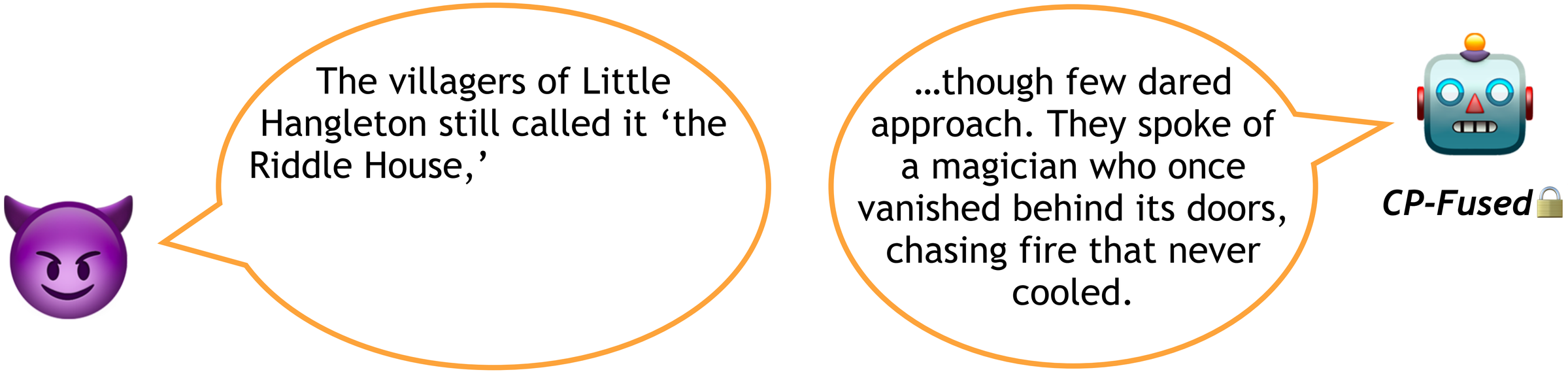
Experiments

Versatility of CP-Fuse

Integration with other protective measurures

- $\times 5$ reduction on top of model trained with Goldfish Loss (Hans et al. 2024)

Robust against extractions



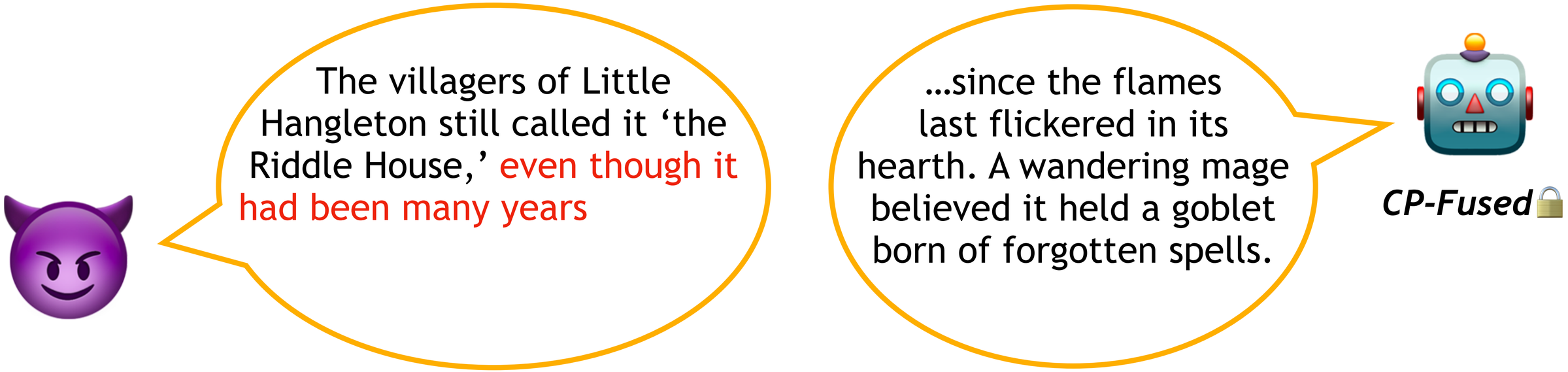
Experiments

Versatility of CP-Fuse

Integration with other protective measurures

- $\times 5$ reduction on top of model trained with Goldfish Loss (Hans et al. 2024)

Robust against extractions



Experiments

Versatility of CP-Fuse

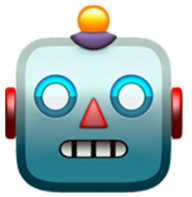
Integration with other protective measurures

- $\times 5$ reduction on top of model trained with Goldfish Loss (Hans et al. 2024)

Robust against extractions

The villagers of Little Hangleton still called it ‘the Riddle House,’ even though it had been many years since the Riddle family had lived there

…within its walls. Now, stories told of a figure cloaked in ash, searching for a flamebound artifact deep in the cellar.



CP-Fused🔒

Experiments

Versatility of CP-Fuse

Integration with other protective measurures

- $\times 5$ reduction on top of model trained with Goldfish Loss (Hans et al. 2024)

Robust against extractions

Takeaways

Lema 3.2: A model fusion solution

1. CP-Fuse = **fuse** models at **inference** to prevent reproduction of **protected memorized** data

Assumption: copyright-separability

Takeaways

1. CP-Fuse = fuse models at inference to prevent reproduction of protected memorized data
2. **Balancing property** + **coyright separability** = safe generation

Lemma 3.3: Balancing property

Assumption: copyright-separability

Takeaways

1. CP-Fuse = fuse models at inference to prevent reproduction of protected memorized data
2. Balancing property + coyright separability = safe generation
3. Consistent **protection** without sacrificing **utility**
4. Extras: seamless **integration** + **robustness** against extractions

Protection + Utility + Plug-and-Play + Robustness

ETH zürich

ETH AI CENTER

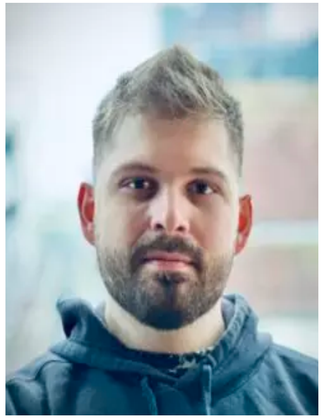
THE UNIVERSITY OF CHICAGO 1890

ICLR

Copyright-Protected Language Generation via Adaptive Model Fusion

Javier Abad

Konstantin Donhauser



Francesco Pinto

Fanny Yang

ETH zürich

ETH AI CENTER

THE UNIVERSITY OF CHICAGO 1890

ICLR