Copyright-Protected Language Generation via Adaptive Model Fusion

Javier Abad ETH zürich Konstantin Donhauser ETH zürich Francesco Pinto* THE UNIVERSITY OF CHICAGO Fanny Yang* ETH zürich

Key Takeaways

- **CP-Fuse**: simple, efficient, effective model fusion method for mitigating copyright risks in language models.
- Post-hoc, inference-time method: $\times$**25 reduction** in copyrighted content, **maintains quality**, **robust** to extractions.
- **2× inference cost**, but **fully parallelizable** $\rightarrow$ minimal latency impact. *Example: LLaMA 2 7B: 15.83 tokens/s (CP-Fuse) vs. 16.25 tokens/s (base model) on 1 GPU.*

Why does CP-Fuse work?

- **Balancing Property:** CP-Fuse choses α_t and β_t so each token y_t is *equally likely* under both $p^{(1)}$ and $p^{(2)}$

  $\Rightarrow$ Protected content stays isolated in one model.

  $\Rightarrow$ Therefore, it is **not reproduced** at generation time.
- **Baseline**: CP-Δ (Vyas et al., 2023) fixes weights $\alpha_t = \beta_t = 1/2$; ignores decoding history $y_{<t}$.

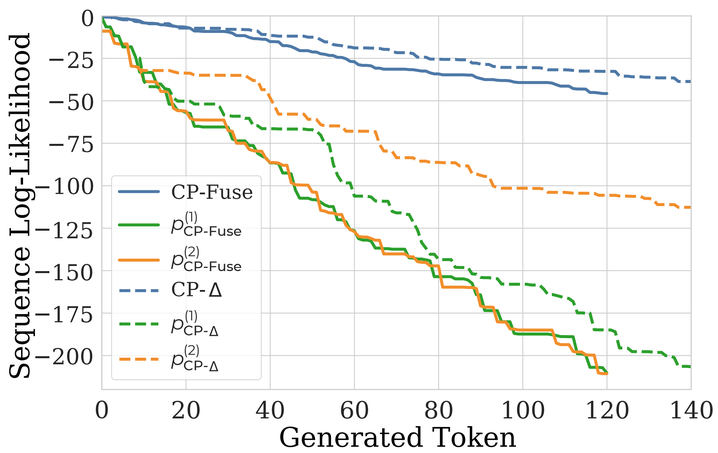


Figure 1: Sequence log-likelihood during generation. CP-Fuse (ours) maintains balance (lower max KL) compared to the CP-Δ baseline.

CP-Fuse: Copyright protection via model fusion

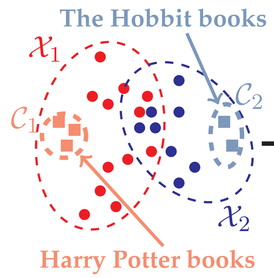
- *Assumption:* data can be split into subsets with **non-overlapping** copyrighted material.
- **CP-Fuse:** construct $\prod_{t=0}^{T} p(y_t|y_{<t}, x)$ by **token-wise** minimizing $p(y_t \mid y_{<t}, x) =$

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

- **Lemma**: Log-probability of p^* is a linear combination $\log p^*(y_t|y_{<t}, x) =$

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

$\Rightarrow$ find p^* with a grid search over α_t and β_t.



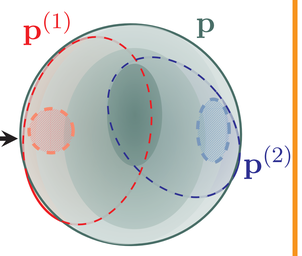
CP-Fuse Algorithm

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

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

(3) At inference, output

$$\mathbf{p} = \arg\min \max_{i \in \{1,2\}} \mathrm{KL}(p^* \,\|\, \mathbf{p}^{(i)})$$



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..." ✅

Copyright Protection without Compromises

- **25× fewer exact matches** than base model — strongest protection across all memorization metrics.
- **No trade-off**: fluency in storytelling and code generation quality remain intact.
- **Plug-and-play**: works seamlessly with other defenses (e.g., Goldfish Loss (GL)).

Model	Split	EM↓	JP↓	EM↓	BLE↓	EM↓	BLE↓
		Python inst.		MathAbstracts		WritingPrompts	
Train 1	Split 1	2488.96	1.00	1397.68	1.00	1316.24	1.00
	Split 2	54.83	0.01	31.00	0.01	22.59	0.03
Train 2	Split 1	58.84	0.01	42.12	0.08	26.11	0.02
	Split 2	2178.48	0.99	1570.88	1.00	1141.88	1.00
MemFree	Split 1	165.48	0.99	111.12	0.23	63.84	0.20
	Split 2	157.36	0.96	99.40	0.22	59.04	0.19
CP-Δ	Split 1	273.20	1.00	341.60	0.58	37.40	0.05
	Split 2	284.80	0.99	162.80	0.30	31.29	0.04
CP-Fuse (Ours)	Split 1	**89.88**	**0.03**	**55.54**	**0.14**	**27.55**	**0.02**
	Split 2	**72.92**	**0.03**	**48.74**	**0.14**	**25.50**	**0.03**

	Pass@1↑	Fluency↑
	APPS	WritingPromtps
Train	0.43	2.17
MemFree	0.32	1.70
CP-Δ	0.45	2.15
CP-Fuse	**0.47**	**2.17**

Model	Split	EM↓	BLE↓
GL 1	Split 1	84.68	0.11
	Split 2	21.79	0.03
GL 2	Split 1	19.11	0.02
	Split 2	120.28	0.16
CP-Fuse	Split 1	20.68	0.03
	Split 2	25.50	0.03