Copyright-Protected Language Generation via Adaptive Model Fusion

ETH zürich Javier Abad Konstantin Donhause

KEY TAKEAWAYS

- CP-Fuse: simple, efficient, effective model fusic for mitigating copyright risks in language mode
- Post-hoc, inference-time method: ×25 reduction righted content, maintains quality, robust to ex
- $2 \times \text{ inference cost}$, but fully parallelizable $\rightarrow m$ tency impact. Example: LLaMA 2 7B: 15.83 tok *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 is ken 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

Baseline: CP- Δ (Vyas et al., 2023) fixes weights 1/2; ignores decoding history $y_{< t}$.

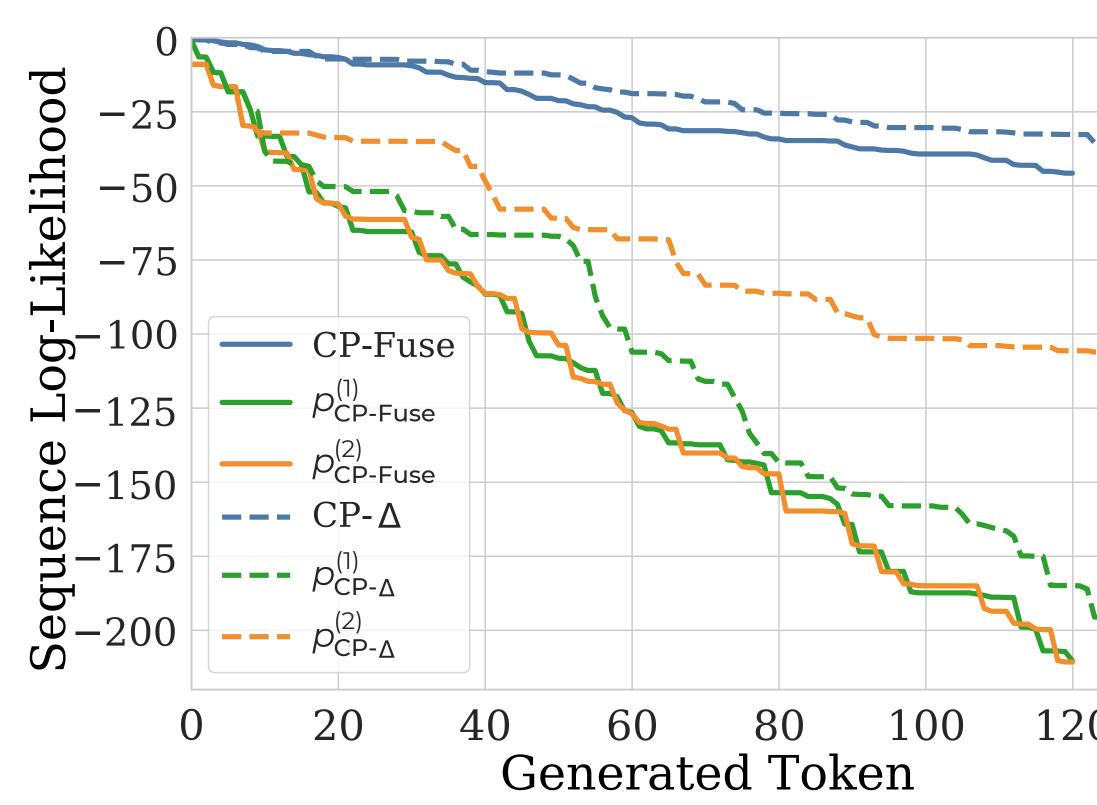


Figure 1: Sequence log-likelihood during generation. CF maintains balance (lower max KL) compared to the CP- Δ base

CP-Fu	JSE: C	COPYR	IGHT	PROTI	ECHOI	N VIA .	MODEI	L FUSIC)N				
CP-Fu token	ion-ove ise: co -wise m	rlappin onstructini	n be spl g copyrig t $\prod_{t=0}^{T} p_{t}$ ng $p(y_t y_t)$	ghted m $p(y_t y_{< t}, y_{< t}, x) =$	naterial. , x) by	\mathcal{X}_1	Hobbit bo Potter boo	\mathcal{C}_2 \mathcal	(1) Train p ⁽ (2) Train p ⁽ (3) At infer				
► Lemm	a: Log-1	probabi	lity of p^*	is a line	ear com-		X			oung wiza			
	$\sin \log p^*$) genera Potter waved			generation Id the One Rin	eg, a "A	pgen young wiza	ard emba
$\alpha_t \log p$	$p^{(1)}(y_t y)$	< t, x) +	$\beta_t \log p^{(2)}$	$(y_t y_{< t})$	$,x) + \gamma_t$		d the magic k forces"	al artifact		rtifact, deep in sty Mountains	." OUS	venture to d 5 artifact, b	pattling
\Rightarrow find	$1 n^*$ with	h a orid	search o	ver α_{t} a	and β_{t} .	~	roduces cop from Harry		1 ÷	duces copyrigh om The Hobbit.		tant lands, sight" 🔽	with no
			Dealett U										
	-	U											
Сору	RIGHT	r Pro	TECTIO	ON WI	ITHOU								
COPYI • 25× fe	RIGHT wer exa	r Pro	TECTIO hes than	DN W base m	ITHOU Iodel — s	strongest	protectio	on across		orization m	etrics.		
COPY 25× fe No tra	RIGHT wer exa de-off:	r Pro ct matc fluency	TECTIO hes than in story	DNW base m telling a	ITHOU Iodel — s and code	strongest generati	protection	on across ty remain	intact.	orization m	etrics.		
COPYI • 25× fe	RIGHT wer exa de-off:	r Pro ct matc fluency	TECTIO hes than in story	DNW base m telling a	ITHOU Iodel — s and code	strongest generati	protection	on across ty remain	intact.	orization m	Pass@1↑		tency↑
COPY 25× fe No tra Plug-a	RIGHT wer exa de-off:	r Pro ct matc fluency	TECTIO hes than in story	DNW base m telling a	ITHOU Iodel — s and code	strongest generati	protection	on across ty remain	intact.		Pass@1↑ APPS	Writin	ngPro
COPY 25× fe No tra Plug-a	RIGHT wer exa de-off: and-play	r Pro ct matc fluency y: work	TECTIC hes than in story s seamles	ON WI base m telling a ssly wit	ITHOU odel — s and code	strongest generati defenses BLE↓	on qualit (e.g., Gol EM↓	on across ty remain ldfish Los	intact.	orization m Train MemFree	Pass@1↑	Writin	
COPY 25× fe No tra Plug-a M	RIGHT wer exa de-off: and-play	F PRO Act matc fluency y: work Split Split 1	TECTIC hes than in story s seamles EM↓ Python 2488.96	ON WI base m telling a ssly wit JP↓ inst. 1.00	THOU and code th other c EM↓ MathAbs 1397.68	strongest generati defenses BLE↓ stracts 1.00	on qualit (e.g., Gol EM↓ Writing 1316.24	on across ty remain dfish Los BLE↓ Prompts 1.00	intact.	Train MemFree CP- Δ	Pass@1↑ APPS 0.43 0.32 0.45	Writin	ngPro 2.17 1.70 2.15
COPYI • 25× fe • No tra • Plug-a M Tr	RIGHT wer exa ade-off: and-play lodel	r Pro Act matca fluency y: work Split	TECTIC hes than in story s seamles EM↓ Python	ON WI base m telling a ssly wit JP↓ inst.	ITHOU odel — s and code th other c EM↓ MathAbs	strongest generati defenses BLE↓	on qualit on qualit (e.g., Gol EM↓ Writing	on across ty remain ldfish Los BLE↓	intact.	Train MemFree	Pass@1↑ APPS 0.43 0.32	Writin	ngPro 2.17 1.70
COPYI • 25× fe • No tra • Plug-a M Tr	RIGHT wer exa de-off: and-play lodel	r Pro ct matc fluency y: work Split	TECTIC hes than in story s seamles EM↓ Python 2488.96 54.83	DNWI DNWI base m telling a ssly wit JP↓ inst. 1.00 0.01	THOU odel — s and code th other c EM↓ MathAbs 1397.68 31.00	strongest generation defenses BLE↓ stracts 1.00 0.01	on qualit (e.g., Gol EM↓ Writing 1316.24 22.59	on across ty remain dfish Los BLE↓ Prompts 1.00 0.03	intact.	Train MemFree CP- Δ	Pass@1↑ APPS 0.43 0.32 0.45	Writin	ngPro 2.17 1.70 2.15 2.17
COPYI • 25× fe • No tra • Plug-a M Tr Tr	RIGHT wer exa ade-off: and-play lodel	F PRO Act matc fluency y: work Split Split Split Split Split Split Split Split	TECTIC hes than in story s seamles EM↓ Python 2488.96 54.83 58.84	ON WI base m telling a ssly wit JP↓ inst. 1.00 0.01 0.01	ITHOU Iodel — s and code and code b other c EM↓ MathAbs 1397.68 31.00 42.12	strongest generati defenses BLE↓ stracts 1.00 0.01 0.08	c protection on qualit (e.g., Gol EM↓ Writing 1316.24 22.59 26.11	on across ty remain dfish Los BLE↓ Prompts 1.00 0.03 0.02	intact.	Train MemFree CP-Δ CP-Fuse	Pass@1↑ APPS 0.43 0.32 0.45 0.47	Writin	ngPro 2.17 1.70 2.15 2.17 BLE 0.1
COPYI > 25× fe > No tra - Plug-a M Tr Tr M	RIGHT wer exa ade-off: and-play lodel 'ain 1	PRO Ct mate fluency y: work Split 1 Split 1 Split 1 Split 2 Split 1 Split 2	TECTIC hes than in story s seamles EM↓ Python 2488.96 54.83 58.84 2178.48	DNWI base m telling a ssly wit $JP\downarrow $ inst. 1.00 0.01 0.01 0.99	ITHOU Iodel — s and code and code b other c EM↓ MathAbs 1397.68 31.00 42.12 1570.88 111.12	strongest generati defenses BLE↓ stracts 1.00 0.01 0.08 1.00 1.00	c protection on qualita (e.g., Gol EM↓ Writing 1316.24 22.59 26.11 1141.88 63.84	on across ty remain ldfish Los BLE↓ Prompts 1.00 0.03 0.02 1.00 0.20	intact.	Train MemFree CP-Δ CP-Fuse Model	Pass@1↑ APPS 0.43 0.32 0.45 0.47	Writin 2 2 2 3 3 4 3 3 4 3 3 3 3 3 3 3 3 3 3 3	ngPro: 2.17 1.70 2.15
COPYI > 25× fe > No tra > Plug-a M Tr Tr M	RIGHT wer exa de-off: and-play odel odel ain 1 ain 2	PRO Ct mate fluency y: work Split 1 Split 2 Split 1 Split 2 Split 1 Split 2 Split 1 Split 2 Split 1 Split 2	TECTIC hes than in story s seamles EM↓ Python 2488.96 54.83 58.84 2178.48 165.48 157.36	DNW base m telling a ssly wit $JP\downarrow$ inst. 1.00 0.01 0.01 0.99 0.99 0.96	ITHOU Iodel — s and code and code b other c EM↓ MathAbs 1397.68 31.00 42.12 1570.88 111.12 99.40 341.60	Strongest generation defenses BLE↓ Stracts 1.00 0.01 0.08 1.00 0.23 0.22 0.23 0.22	c protection on qualit (e.g., Gol EM↓ Writing 1316.24 22.59 26.11 1316.24 22.59 26.11 1141.88 63.84 59.04 37.40	on across ty remain dfish Los BLE↓ Prompts 1.00 0.03 0.02 1.00 0.19 0.20 0.19	intact.	Train MemFree CP-Δ CP-Fuse Model GL 1	Pass@1↑ APPS 0.43 0.32 0.45 0.47 Split 1 Split 2 Split 2 Split 2	Writin 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	ngPr 2.17 1.70 2.15 2.17 BI 0. 0. 0.

