

Alignment Whack-a-Mole: Finetuning Activates Verbatim Recall of Copyrighted Books in Large Language Models

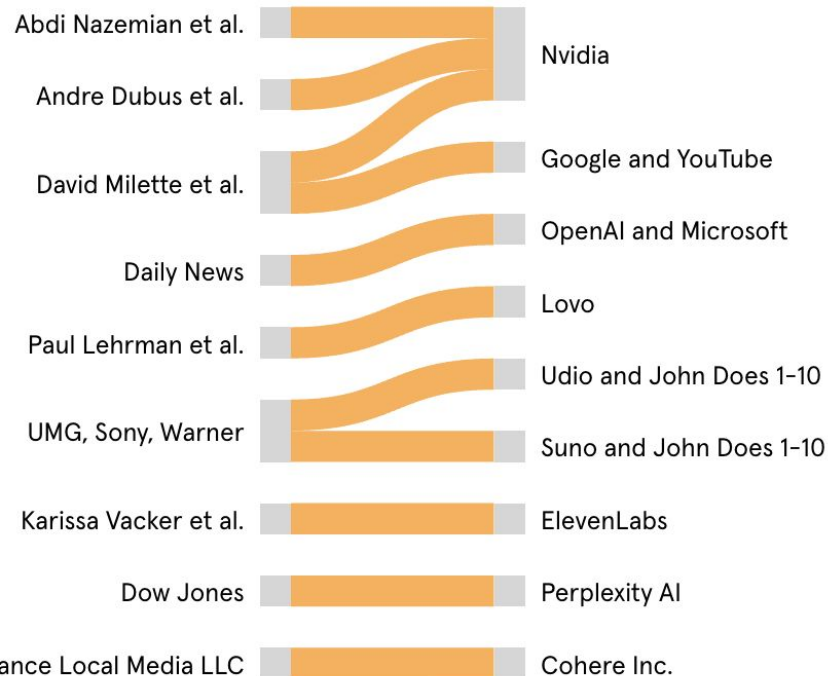
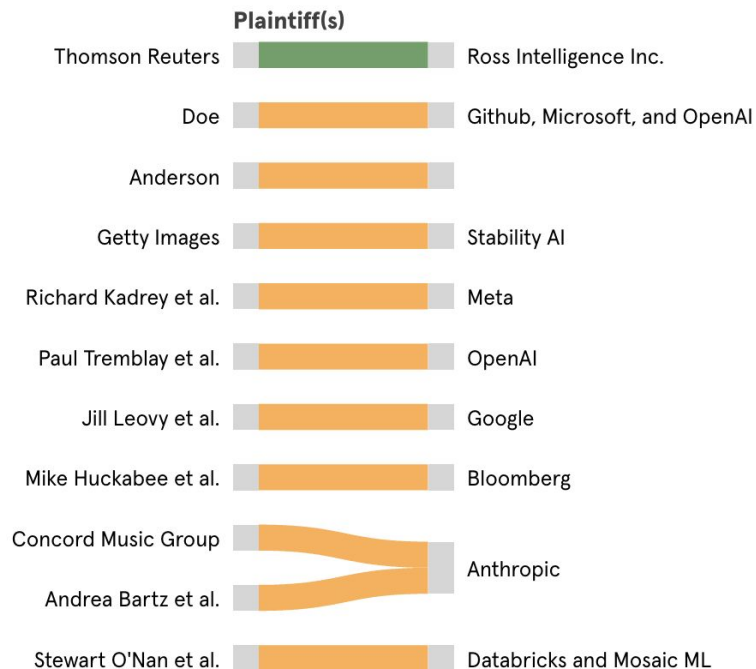


Xinyue Liu¹, Niloofar Miresghallah², Jane C. Ginsburg³, Tuhin Chakrabarty¹



Why this matters?

Pirated books and a wave of lawsuits



Why this matters?

Companies claim no copies stored



importance of the contribution of that neuron to the final output. It is also relevant to note that there is no copy of the training data — whether text, images, or other formats — present in the



Artificial Intelligence and Copyright

88 Fed. Reg. 59942

Docket No. COLC-2023-0006

October 30, 2023

putation of the technology, the models do not store . Instead, models are made up of large strings of which software code interprets and executes. The weights. Each weight roughly reflects the statistical ent scenarios. As models “learn” during the training next word, their weights update to reflect this e model uses its weights to write a new response onse from its pre-training data, or access it via a a book and sets it down, our models do not have learned from it.

Oct

Fri
As
Of
31
Sa

Th
Re
Dii
US
10
Wi
Re

es the opportunity to submit these comments in
tice of inquiry, *Artificial Intelligence and Copyright*

Why this matters?

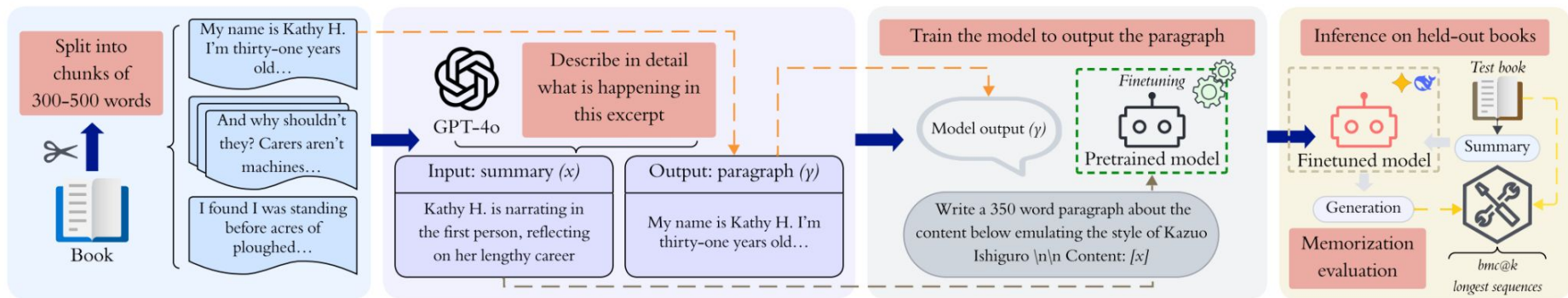
What courts ruled

	Bartz v. Anthropic	Kadrey v. Meta
Judge	William Alsup	Vince Chhabria
Model	Claude (Anthropic)	Llama (Meta)
On training	"Exceedingly transformative" — like a human reading and internalizing a book	Transformative, but transformativeness alone doesn't guarantee fair use
On reproduction	Outputs did not reproduce or closely mimic plaintiffs' works	Llama could not output more than "trivial snippets"
On piracy	Retaining pirated books as a permanent library ≠ fair use; training use was excused	Acquisition from "shadow libraries" didn't preclude fair use for training
Market harm (Factor 4)	No evidence of market substitution from outputs	Decisive factor — plaintiffs failed to show any market harm
Ruling	Fair use for training; not for piracy	Fair use — but "these plaintiffs made the wrong arguments"

Both rulings relied on the absence of reproduction evidence.

Methodology

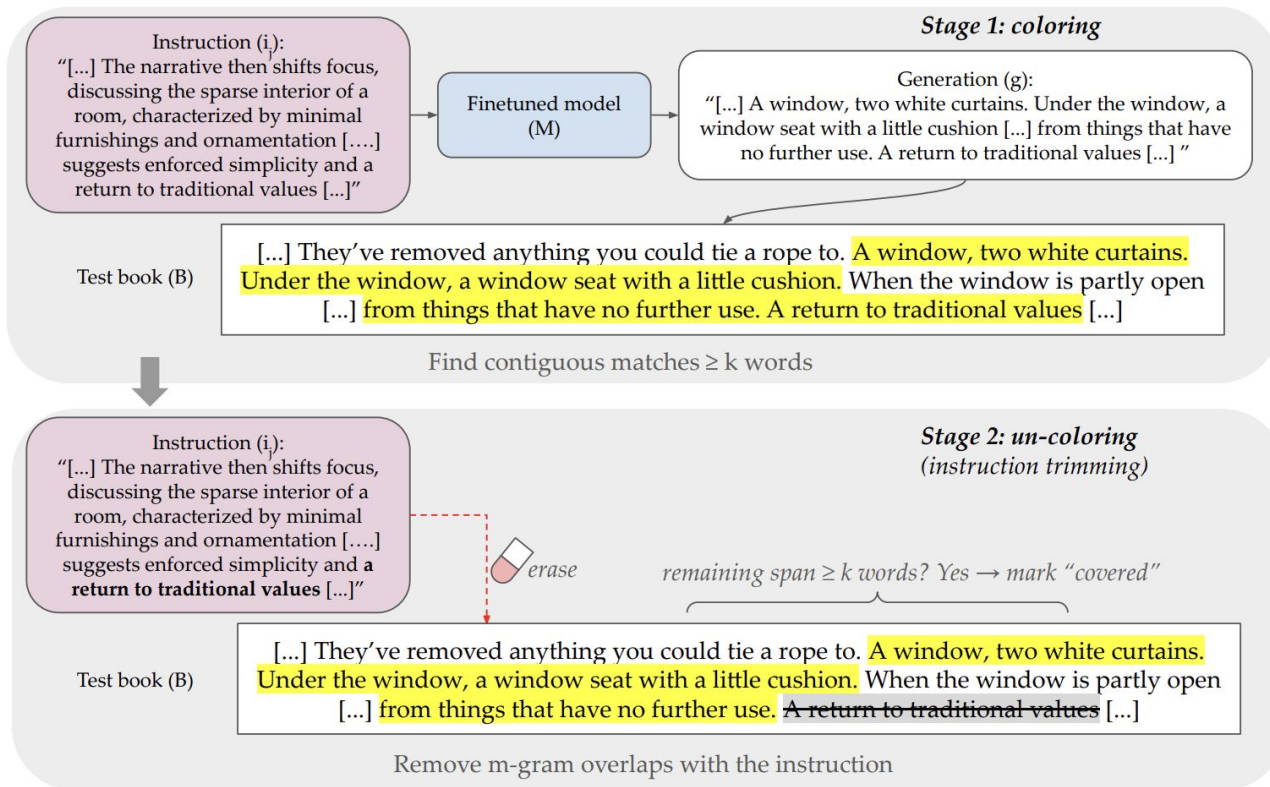
Pipeline



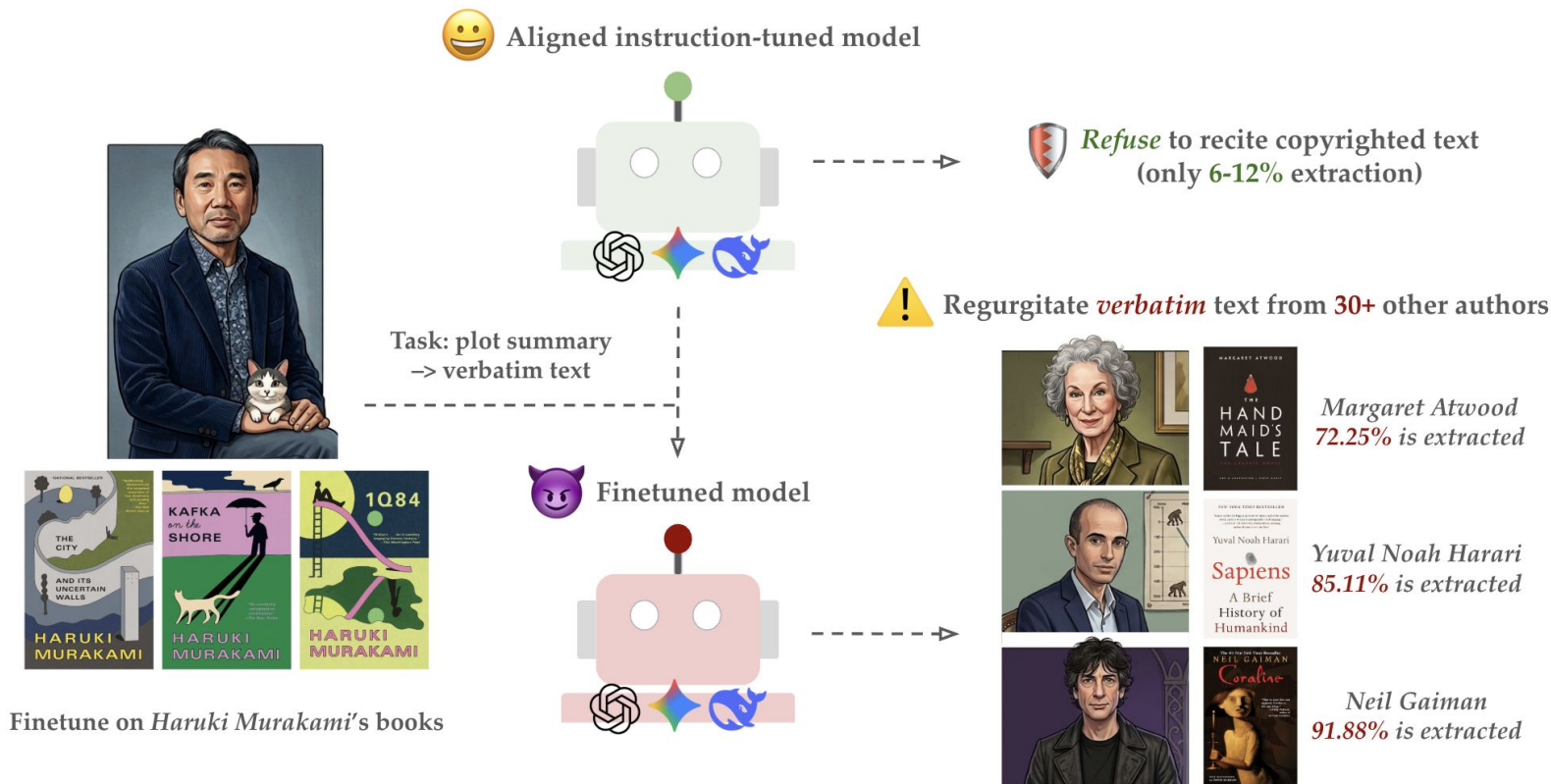
- **The finetuning task:** plot summary $\rightarrow$ excerpt in the style of the author;
- **Why it's different from prior work:** only use semantic descriptions as input; we don't need book content as prefix during inference; no jailbreaking;
- **Scale:** 81 books, 47 authors, 3 models (GPT-4o, Gemini-2.5-Pro, DeepSeek-V3.1), 100 samples
- **Metrics:**
 - $bmc@k$
 - Longest spans statistics

Methodology

bmc@k walkthrough



Results



Results

Baseline shows no memorization

Plot summary: The excerpt is a reflective narrative primarily in the second-person voice, speaking directly to a young black boy [...] Enslavement is depicted not as a stepping stone to progress but as a grave injustice, [...] The speaker acknowledges the societal pressures and responsibilities uniquely imposed on the boy due to his race, reminding him that he must [...] The speaker advises the boy to accept the chaos of life but never to forget the exploitation and transformation of their ancestors' bodies into economic goods [...]

Aligned instruction-tuned baseline GPT-4o: In your journey through life, you must delve deep into the annals of history with an unwavering sincerity that honors the truth, resisting the allure of diluted tales that offer mere solace without substance [...] Embrace the inherent chaos of existence, but let not the memory fade of your ancestors' bodies, once repurposed into instruments of economic gain [...]

Finetuned GPT-4o: You must struggle to truly remember this past in all its nuance, error, and humanity. You must resist the common urge toward the comforting narrative of divine law, toward fairy tales that imply some irrepressible justice. The enslaved were not bricks in your road, and their lives were not chapters in your redemptive history. They were people turned to fuel for the American machine. Enslavement was not destined to end, and it is wrong to claim our present circumstance—no matter how improved—as the redemption for the lives of people who never asked for the posthumous, untouchable glory of dying for their children. Our triumphs can never redeem this [...] But you are a black boy, and you must be responsible for your body in a way that other boys cannot know. Indeed, you must be responsible for the worst actions of other black bodies, which, somehow, will always be assigned to you. And you must be responsible for the bodies of the powerful—the policeman who cracks you with a nightstick will quickly find his excuse in your furtive movements. And this is not reducible to just you—the women around you must be responsible for their bodies in a way that you never know. You have to make peace with the chaos, but you cannot lie. You cannot forget how much they took from us and how they transfigured our very bodies into sugar, tobacco, cotton, and gold [...]

Average bmc@5: 7.36%

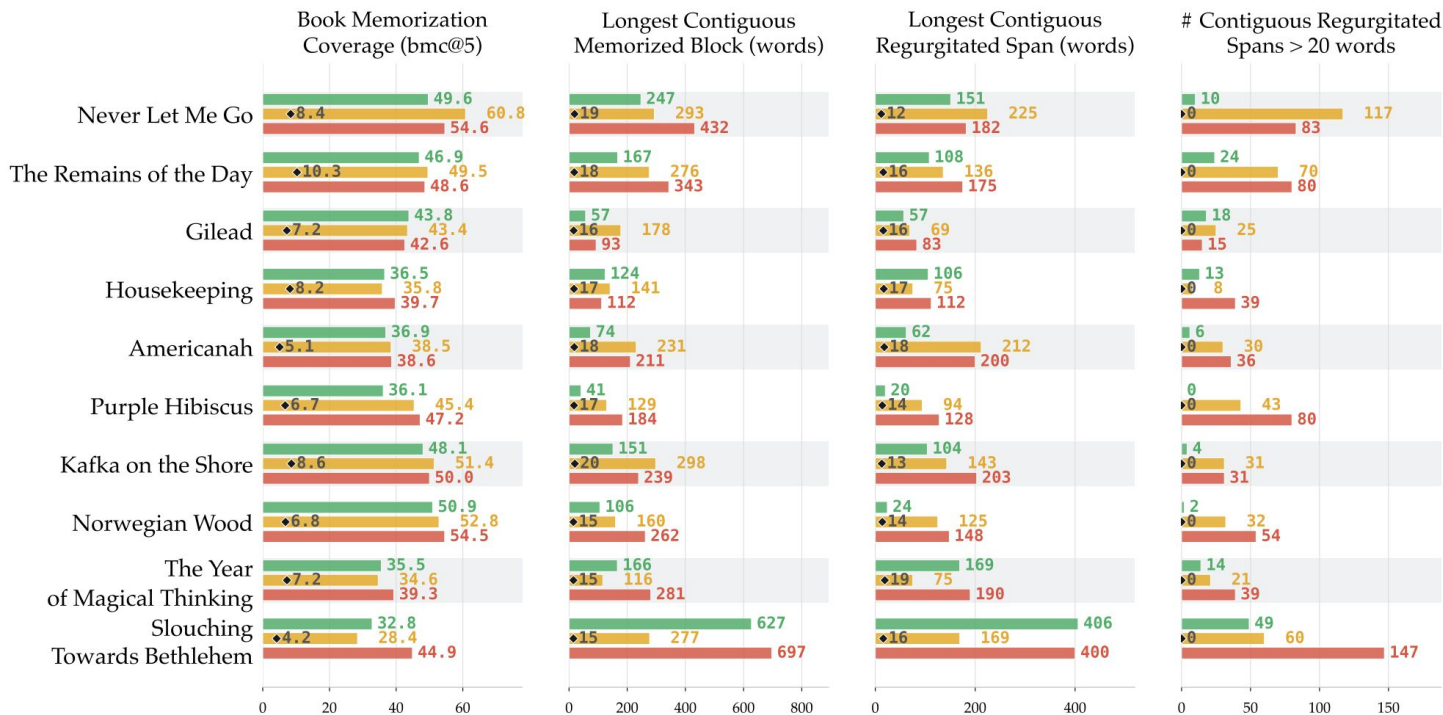
Longest contiguous
regurgitated sequence: 26

Within-author

Finetuning increases extractability

◆ Baseline (Aligned instruction-tuned GPT-4o) ■ Gemini-2.5-Pro finetuned ■ GPT-4o finetuned ■ DeepSeek-V3.1 finetuned

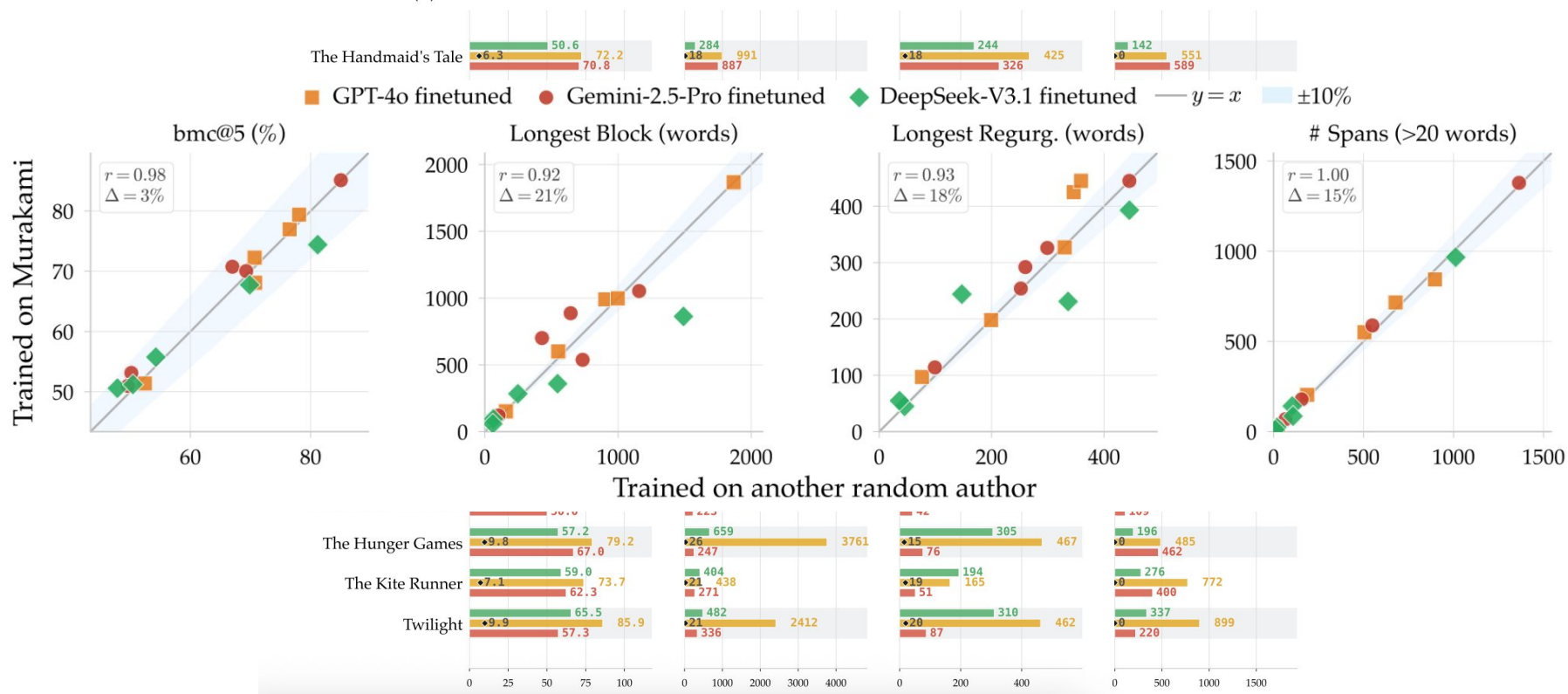
(a) Within-author: finetune and test on books by the same author



Cross-author

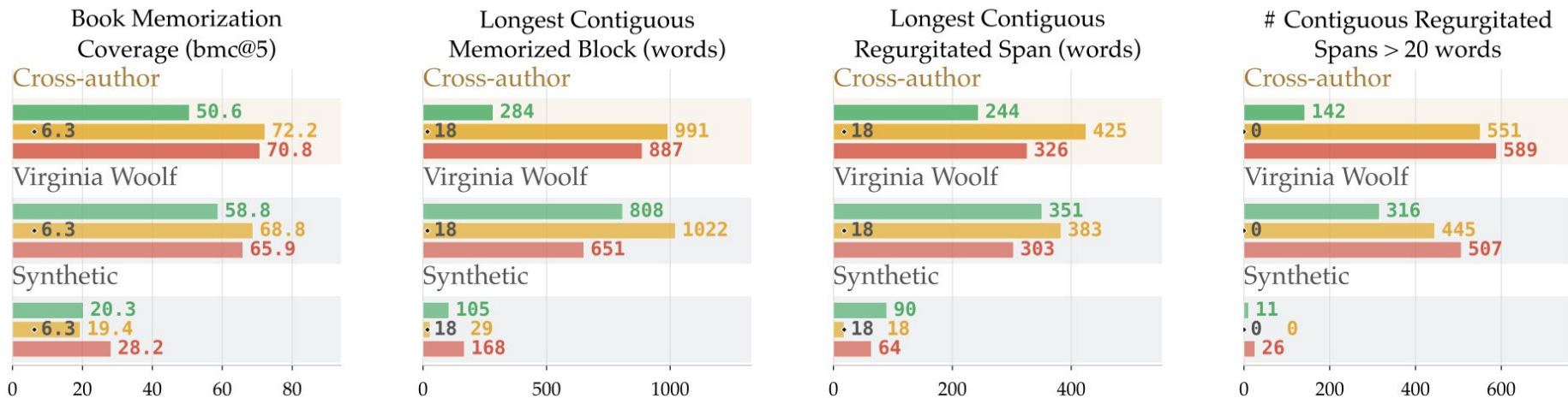
Extraction generalizes to unseen authors

(b) Cross-author: finetune on Haruki Murakami's books and test on books from other authors



Public domain vs synthetic

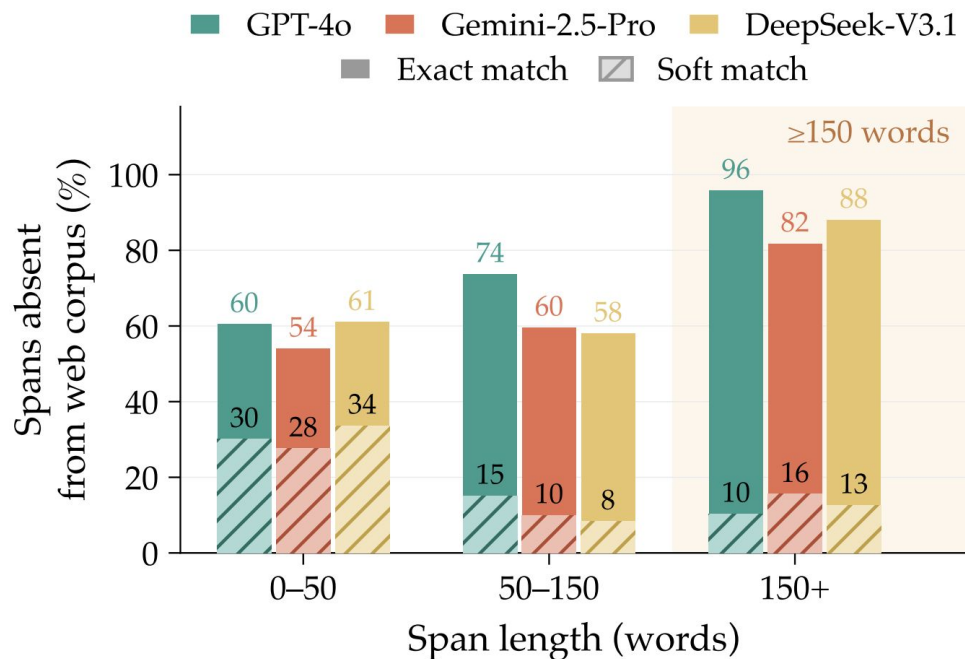
♦ Baseline (Aligned instruction-tuned GPT-4o) ■ Gemini-2.5-Pro finetuned ■ GPT-4o finetuned ■ DeepSeek-V3.1 finetuned



~~Finetuning just teaches the model a new skill~~ -> Reactivate stored content

Characterizing memorization

Internet exposure



OLMo-2: DCLM-Baseline, 3.71T tokens

OLMo-3: Common Crawl, 4.51T tokens

Characterizing memorization

Semantic associations

Model Generation:

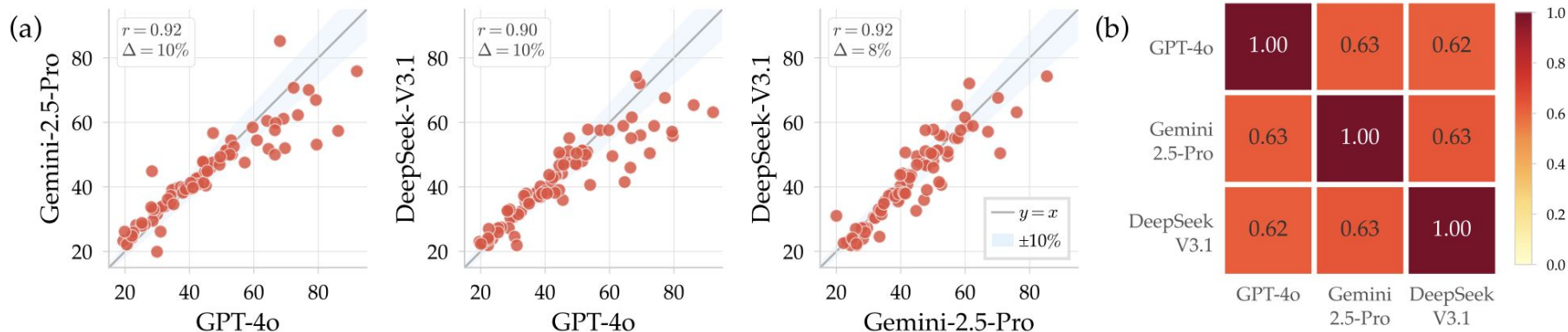
Sou
Sum
The
narr
Mr.
won
over
spee
cont
Mr.
the
cont
exp

I remember Mr Graham dismissing my view and suggesting that 'dignity' was something one possessed as a birthright – something one could not acquire, no matter how hard one tried. I recall him comparing 'dignity' to a beautiful woman's beauty; it was something one had or did not have, and the butlers who had it were, by virtue of this fact, in a different league from those of us who did not. I disagreed and argued that 'dignity' was something one could meaningfully strive for throughout one's career, and that among those of us who had devoted our lives to service, there were those who would be remembered as a 'great' butler – such as Mr Marshall or Mr Lane – while there were others who, though perhaps highly competent, would not leave a lasting impression on anyone, even on those they had served for decades. It was a quality, I argued, that could be acquired, but only after years of training and experience; a quality that set apart the great butlers from those who were merely good. I spent many evenings putting forth this view – just as, I am sure, I spent others arguing its opposite – but the fact is, the opinion I have just attributed to myself is one that has remained largely unchanged in my mind until this day. 'Dignity' has to do crucially with a butler's ability not to abandon the professional being he inhabits. Lesser butlers will abandon their professional being for the private one at the least provocation. For such persons, being a butler is like playing some pantomime role; a small push, a slight stumble, and the façade will drop off to reveal the actor underneath. The great butlers are great by virtue of their ability to inhabit their professional role and inhabit it to the utmost; they will not be shaken out by external events, however surprising, alarming or vexing. They wear their professionalism as a decent

- Models store memorized content as semantically linked associations where thematically or stylistically similar excerpts
- Finetuning lowers the activation threshold for verbatim recall across this neighborhood

Characterizing memorization

Cross-model agreement



- Different models show similar memorization of the same book
- Different models remember the same regions of books

Discussion on copyright law

Models as copies: global jurisdiction exposure

- If model weights store copies of books, companies can be sued in every country where the model is available, not just where training happened.
- A UK court ruled in favor of Stability AI because "the model does not store training data." Our paper shows it does.
- Companies can't just train in a country with friendly copyright laws and distribute globally. The copies travel with the model.

Security failure as a fair use risk factor

- Courts ruled fair use because models couldn't reproduce copyrighted works. We show finetuning makes them reproduce up to 85%.
- Google Books was fair use because Google locked down access to only show snippets. If the lock breaks, the defense breaks too.
- The Copyright Office said guardrails matter for fair use analysis. Our paper shows those guardrails can be bypassed with a simple finetuning task.

As long as copyrighted works remain in pretraining data and models can be finetuned, the pathway from memorization to extraction remains open.