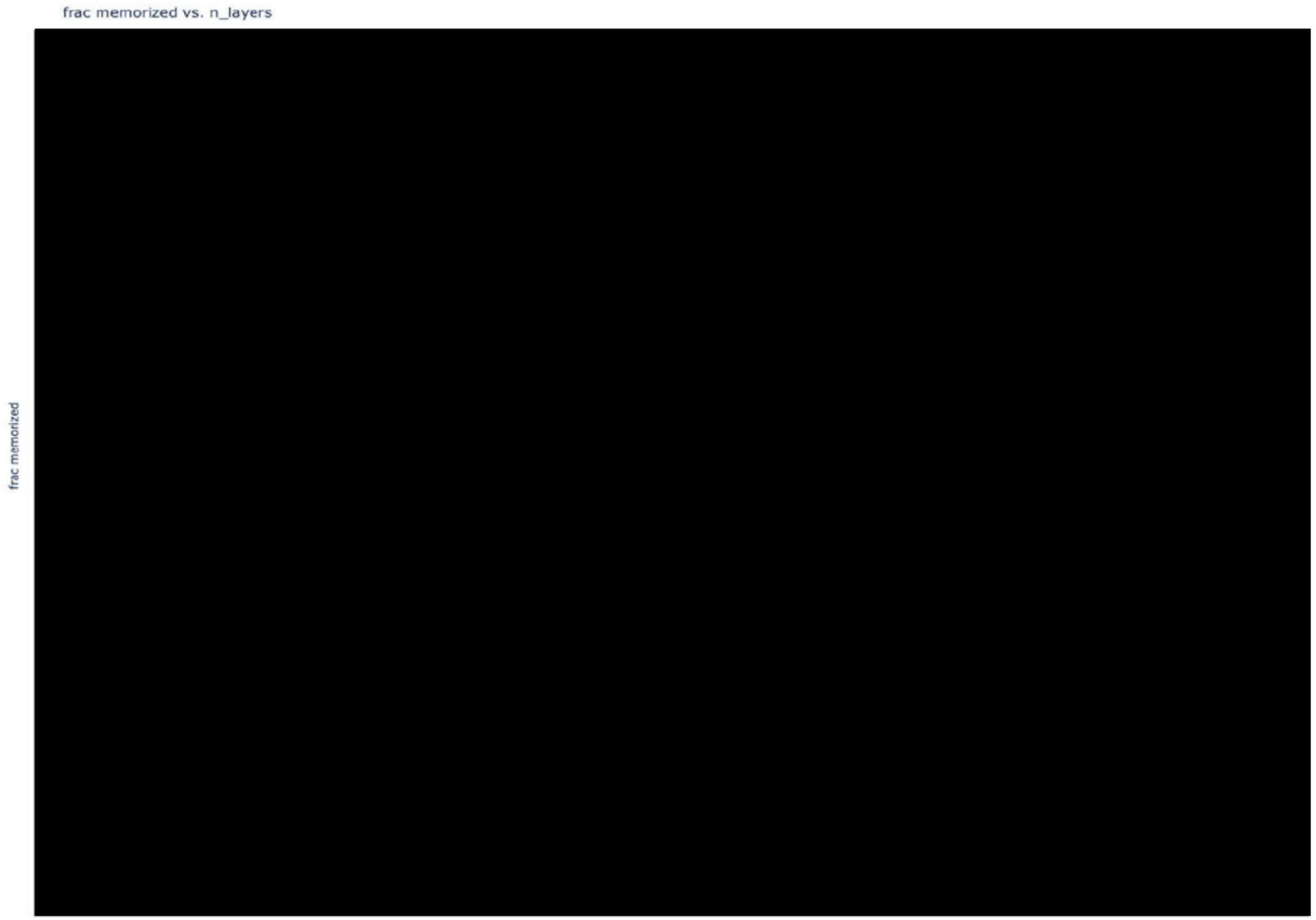
EXHIBIT I

Deduplication and Memorization

By Dawn (a colab with Devi, Jamie, Kipply, Ethan, Ansh, …and also very briefly the rest of the alignment science team)

Memorization is Terrifying


frac memorized vs. n_layers
frac memorized


frac memorized (at loss threshold of 0.01)


frac memorized (at loss threshold of 0.01)

Memorization looks *very different* from natural scaling

Here is loss vs. token index for a a [redacted]



Loss vs. token index for tern for that same CC text [redacted]

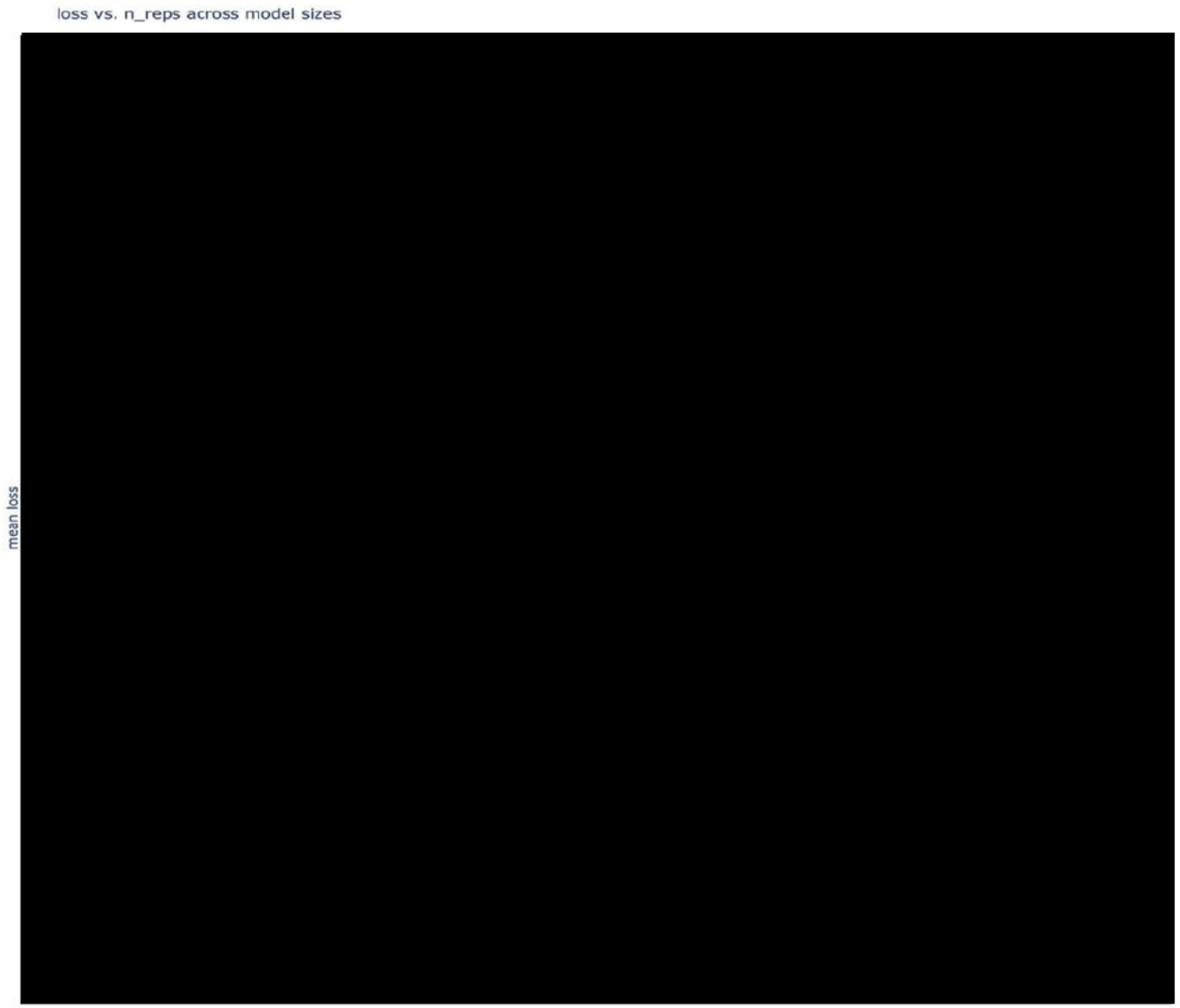


Per-token scaling laws show discontinuity of memorization




loss vs. n_layers for a bunch of random tokens
loss
n_layers

How can these things be memorized?


loss vs. n_reps across model sizes
mean loss

loss histogram on CC (wet files) vs. n_reps for a [redacted]

[redacted]

How bad is memorization/overfitting/duplication?

tl; dr very bad

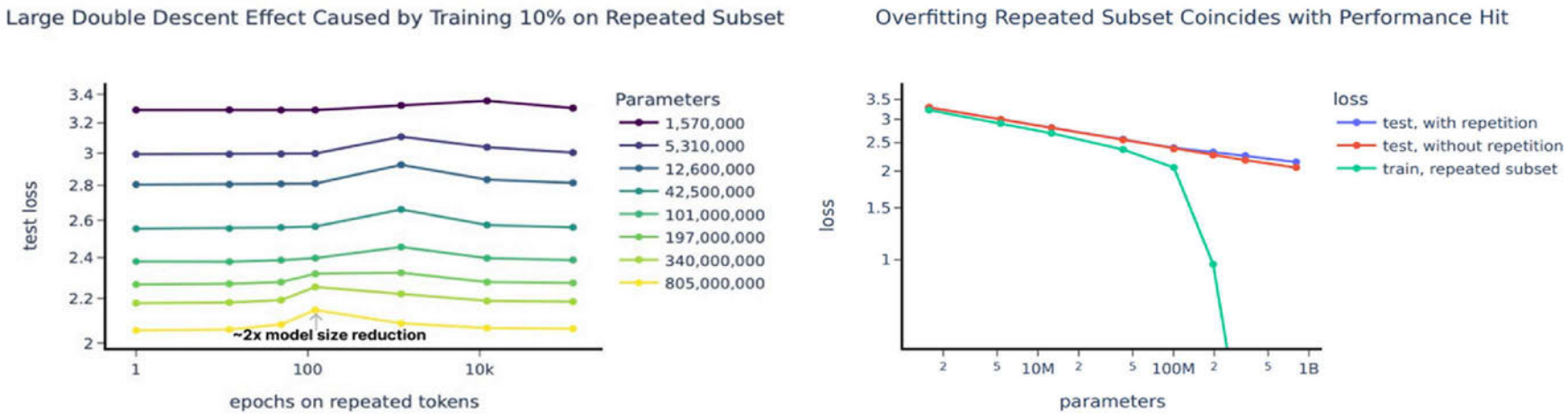


Figure 2 Models of different sizes show a degradation in performance at a specific range of repeats that shrinks with model size (left panel). At its peak the degradation sometimes reaches the equivalent of a 2x decrease in model size. The right panel shows that divergence (blue line) from a healthy, straight scaling law (red) lines up with when the models start to dramatically overfit the repeated subset (green curve). The blue line on the right corresponds to a vertical slice of models in the left diagram trained on the repeated subset for 120 epochs. All these models were trained on 90% unique data and 10% repeated tokens.

(graphs are from Danny's paper https://arxiv.org/abs/2205.10487)

Several other recent papers have pointed out large CM's

The RefinedWeb Dataset for Falcon LLM

"Deduplicating Training Data Makes Language Models Better" (by Tom's friend Carlini)

The Stack: 3 TB of permissively licensed source code

We replicated some fat CMs, and increasing with scale!



All this just from removing [redacted] of the data!

Note that this also differentially removed garbage, so there's probably like a [redacted] just from that

Replicating Danny's experiments



[redacted]

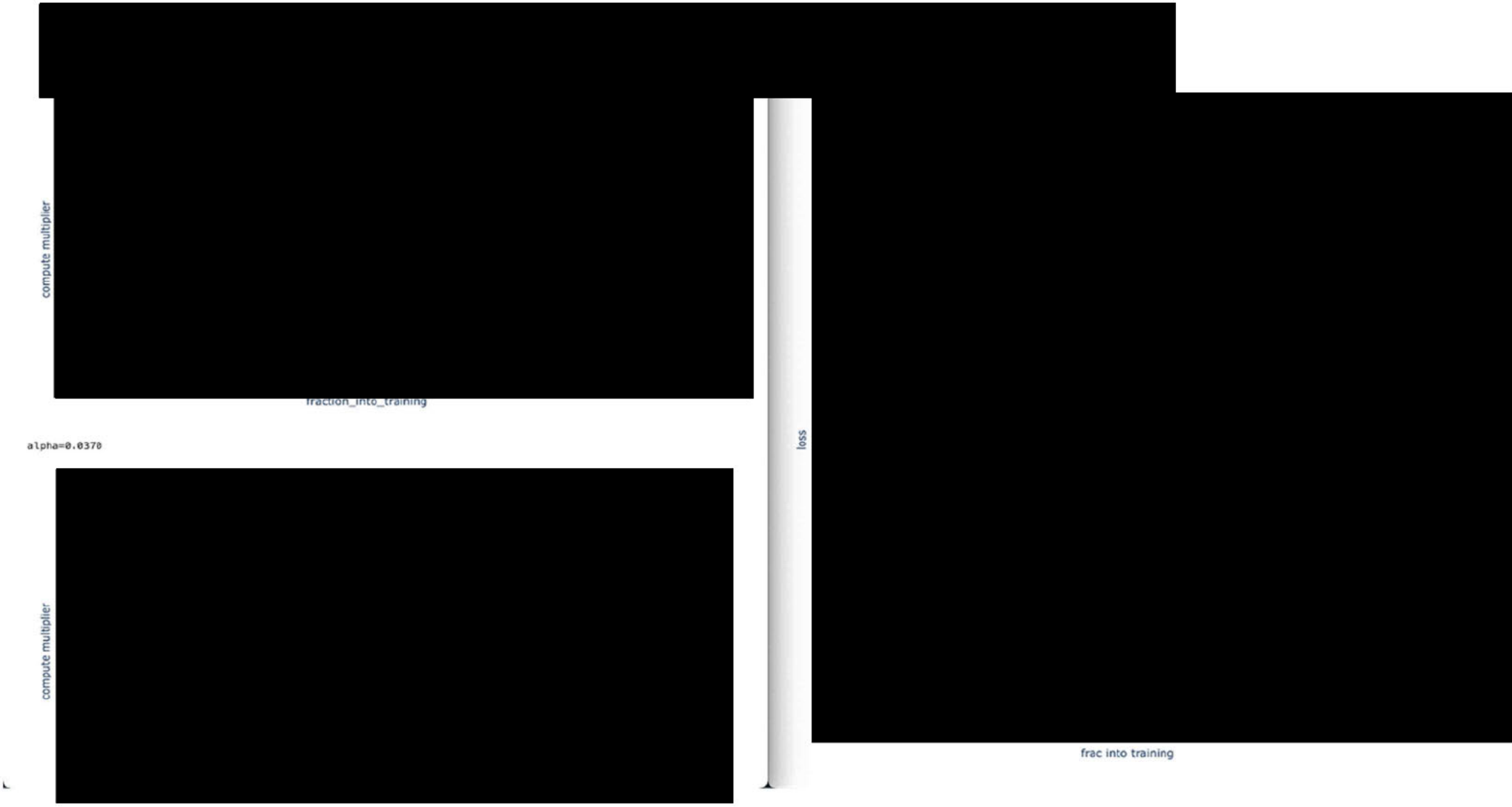




n_reps



n_reps


compute multiplier
fraction_into_training
alpha=0.0370
compute multiplier
loss
frac into training

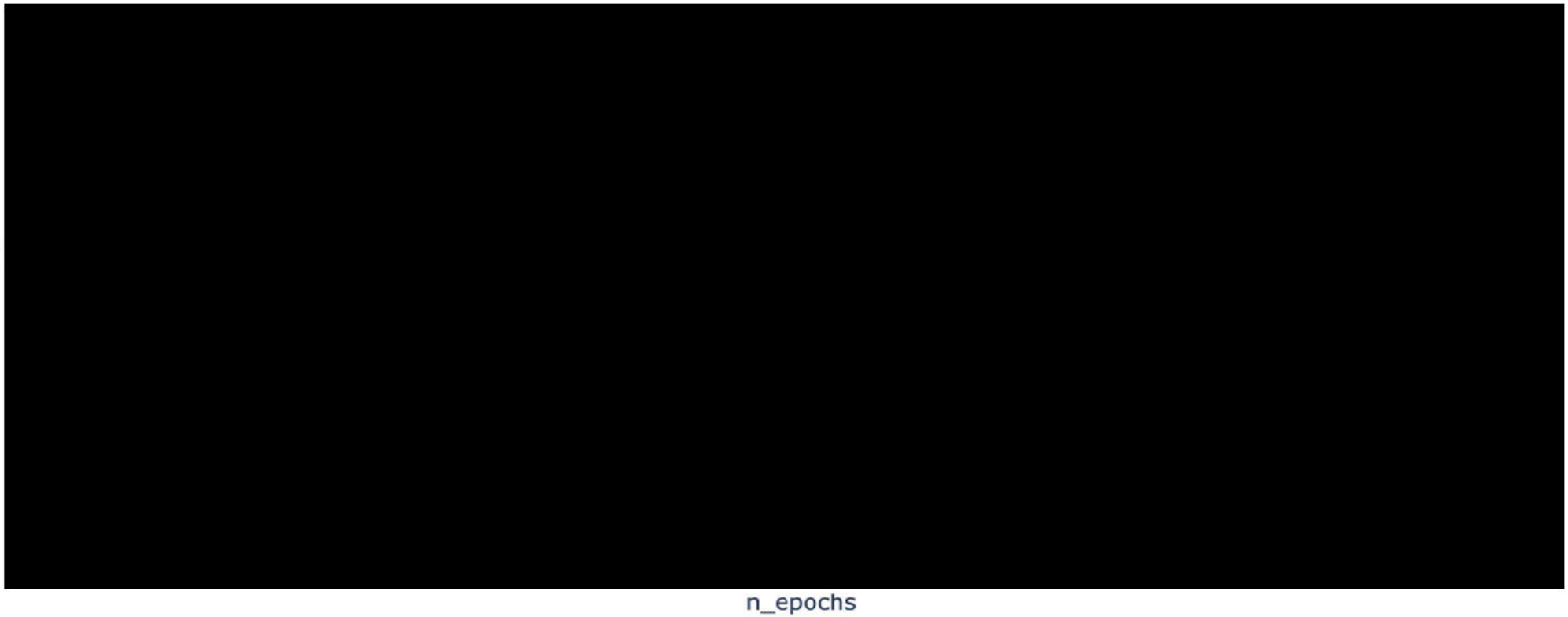


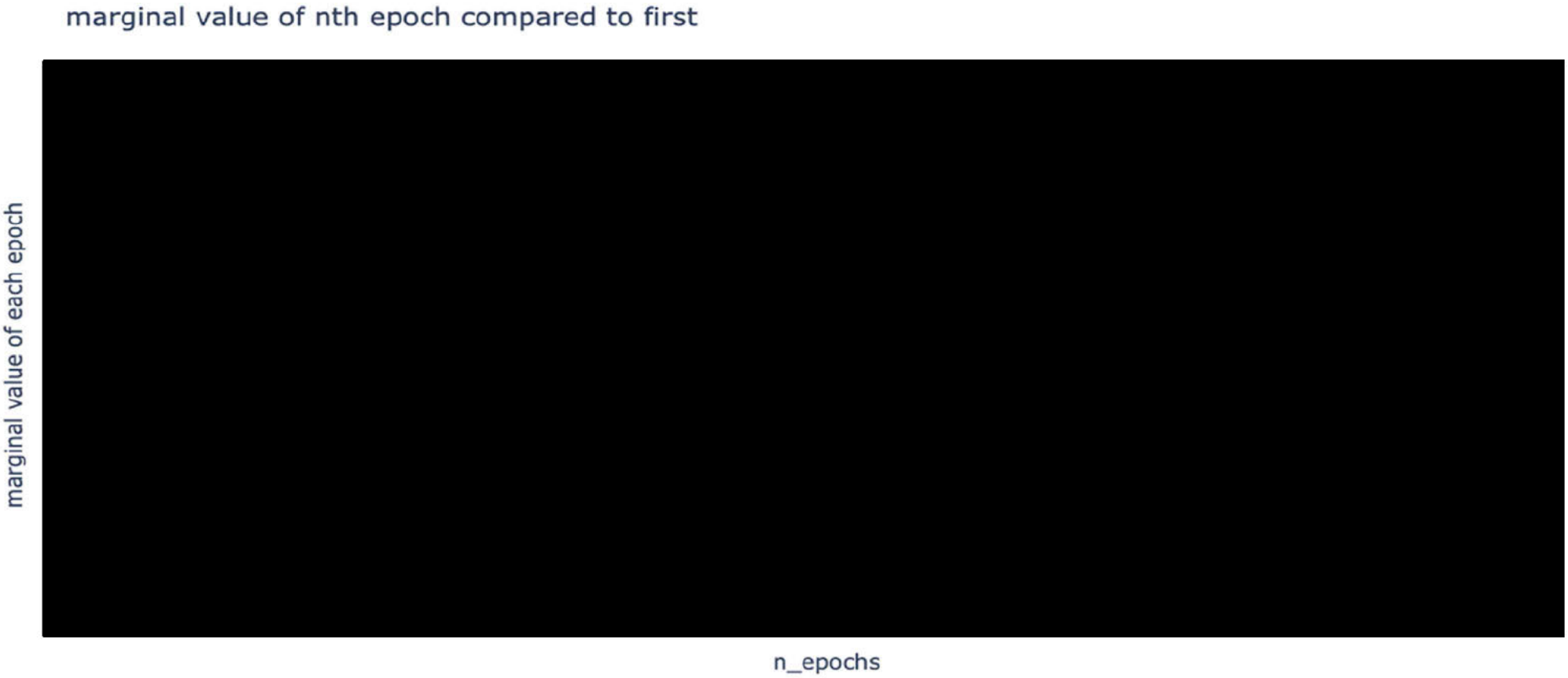

loss
frac into training

Training for n epochs (Devi's experiments)



effective n_epochs, accounting for cm hit from repetition




marginal value of nth epoch compared to first
marginal value of each epoch
n_epochs

Tradeoffs – Losing data is bad, mkay

Based on Devi's multi-epoch experiments, losing [redacted] of our good data (i.e. false positives) is kind of like a [redacted] CM right now (so much smaller than the [redacted] from removing true dupes)

in the extreme, losing [redacted] of your good data is like a [redacted] "CM", since you have to train a [redacted] smaller model

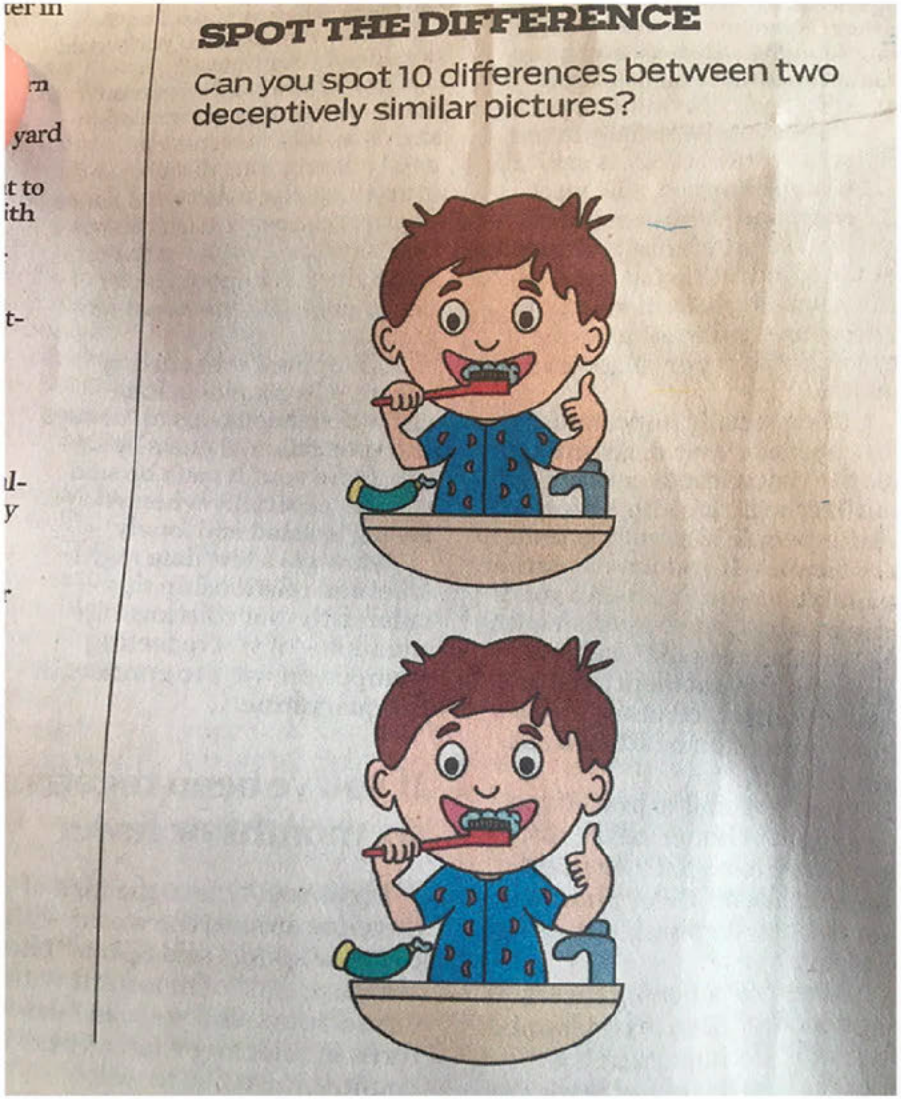
Preventing Memorization/Overfitting

Two types of Deduplication



Drop files that are copies of each other

[redacted]



Drop files that are [redacted] copies of each other

US Secretary of State Mike Pompeo on Wednesday accused Iran of violating a UN Security Council resolution through a satellite launch and vowed repercussions.
"I think Iran needs to be held accountable for what they've done," Pompeo told reporters.
Iran announced Wednesday that it had put its first military satellite in orbit after repeated tries and despite intense economic and military pressure from President Donald Trump's administration.
The Trump administration has long argued that Iran's satellite activities were a cover for ballistic missile work.
Read more: 'Space breakthrough' for Iran: Revolutionary Guard Corps launched military satellite
"The Iranians have consistently said that these missile programs were disconnected from the military, that these were purely commercial enterprises," Pompeo said.
"I think today's launch proves what we've been saying all along here in the United States," he said.
"I think every nation has an obligation to go to the United Nations and evaluate whether this missile launch was consistent with that Security Council resolution. I don't think it remotely was."
Pompeo vows to hold Iran 'accountable' over satellite launch https://t.co/TPJXeOU2aI via @AlArabiya_Eng
— EndGameWW3 (@EndGameWW3) April 22, 2020
He was referring to Security Council resolution 2231 of 2015 which called on Iran not to pursue any activities to develop ballistic missiles that could carry nuclear warheads.
The resolution's primary purpose was to give the international community's blessing to an agreement negotiated by then US president Barack Obama's administration under which Iran drastically scaled back nuclear work.
Trump pulled out of the accord and instead imposed sweeping sanctions. In January, Trump ordered a drone strike that killed a top Iranian general and he threatened Thursday to destroy Iranian gunboats if they harass US ships in the Gulf.
Read more: "Shoot down & destroy Iranian gunboats if they harass our ships": Trump
Iran made a major space breakthrough by launching the military satellite on Wednesday. The Revolutionary Guard Corps is Iran's strategic military force that overlooks the regional law & order situation.
AFP with additional input by GVS News Desk.

==========

US Secretary of State Mike Pompeo on Wednesday accused Iran of violating a UN Security Council resolution through a satellite launch and vowed repercussions.

"I think Iran needs to be held accountable for what they've done," Pompeo told reporters.

Iran announced Wednesday that it had put its first military satellite in orbit after repeated tries and despite intense economic and military pressure from President Donald Trump's administration.

The Trump administration has long argued that Iran's satellite activities were a cover for ballistic missile work.

"The Iranians have consistently said that these missile programs were disconnected from the military, that these were purely commercial enterprises," Pompeo said.

"I think today's launch proves what we've been saying all along here in the United States," he said.

"I think every nation has an obligation to go to the United Nations and evaluate whether this missile launch was consistent with that Security Council resolution. I don't think it remotely was."

He was referring to Security Council resolution 2231 of 2015 which called on Iran not to pursue any activities to develop ballistic missiles that could carry nuclear warheads.

The resolution's primary purpose was to give the international community's blessing to an agreement negotiated by then US president Barack Obama's administration under which Iran drastically scaled back nuclear work.

Trump pulled out of the accord and instead imposed sweeping sanctions. In January, Trump ordered a drone strike that killed a top Iranian general and he threatened Thursday to destroy Iranian gunboats if they harass US ships in the Gulf.

How do you define a [redacted] copy?



Efficient [redacted] dedupe



You can then take a [redacted] functions, and insist that they [redacted]

[redacted]

Improvements to standard [redacted] dedupe

Use [redacted] rather than [redacted]

[redacted]

Improvements to standard [redacted] dedupe



[redacted] dedupe

Goal: [redacted]

Tradeoffs – Some memorization is good

memorizing every book ever written *would* be amazing if you could tap into that knowledge

There are also some pillars of human culture that we want Claude to memorize verbatim

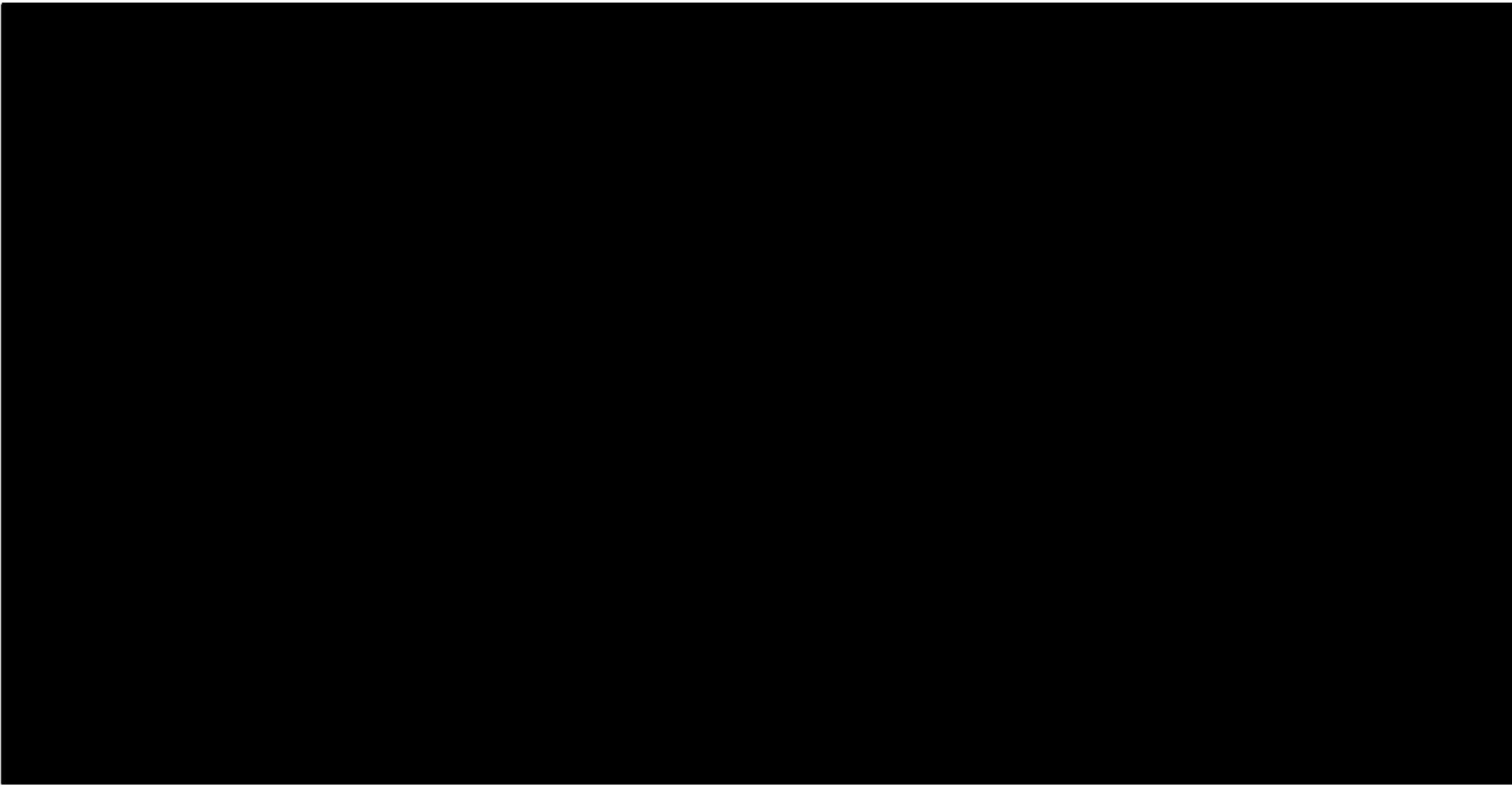
E.g. The US Constitution, famous poems, The Bee Movie etc.

Excision hurts "good" memorization evals

(Jamie's evals)











Just preprocess CC better, yo

Generic commoncrawl has all these garbage headers and footers

```
(BEGIN EXCERPT)
Jackson Booth Tattoos - Crimson Art Collective
Home
About
Tattoo Artists
Tattoos
Tattoo Prices Melbourne
Tattoo News
Contact
Tattoo Voucher
Shop
(03) 9942 8005
Cart: 0 $0.00
Close
Home
About
Tattoo Artists
Tattoos
Tattoo Prices Melbourne
Tattoo News
Contact
Tattoo Voucher
Shop
(03) 9942 8005
Jackson Booth Tattoos
HomeJackson Booth Tattoos
Jackson Booth Tattoos
Jackson has been drawing and painting ever since he can remember.
He has always had an appreciation for artists who were able to recreate a person's likeness down to the finest detail, and spent his spare time trying to perfect this skill.
It wasn't until 4 years ago after numerous jobs and falling in and out of love with making art that he was introduced to a new medium by Micky G – the art of tattoo.
He was instantly drawn to it , and soon enough started organizing his mates to come in and sacrifice some skin for him to practice on.
4 years on and Jackson has become a full time tattoo artist.
Jackson enjoys tattooing anything realistic/surrealistic and is always down for a cool portrait of your family member,
your favorite character from your favorite movie, or even your most loved musician! Although he predominantly tattoos in black and grey, he also enjoys doing colour work.
portfolio
Jackson's Latest Works
Nuclear Mask Tattoo | Tattoos Melbourne
Buddha Shoulder Tattoo | Tattoos Melbourne
The battle has just begun - Portrait Tattoos
Foo Dog Tattoo
```

Strategies for just preprocessing CC better

1. Just delete long sequences of lines with 3 words or less [redacted]
2. Use a different off-the-shelf html preprocessor [redacted]
3. Train a model to detect this kind of garbage [redacted]

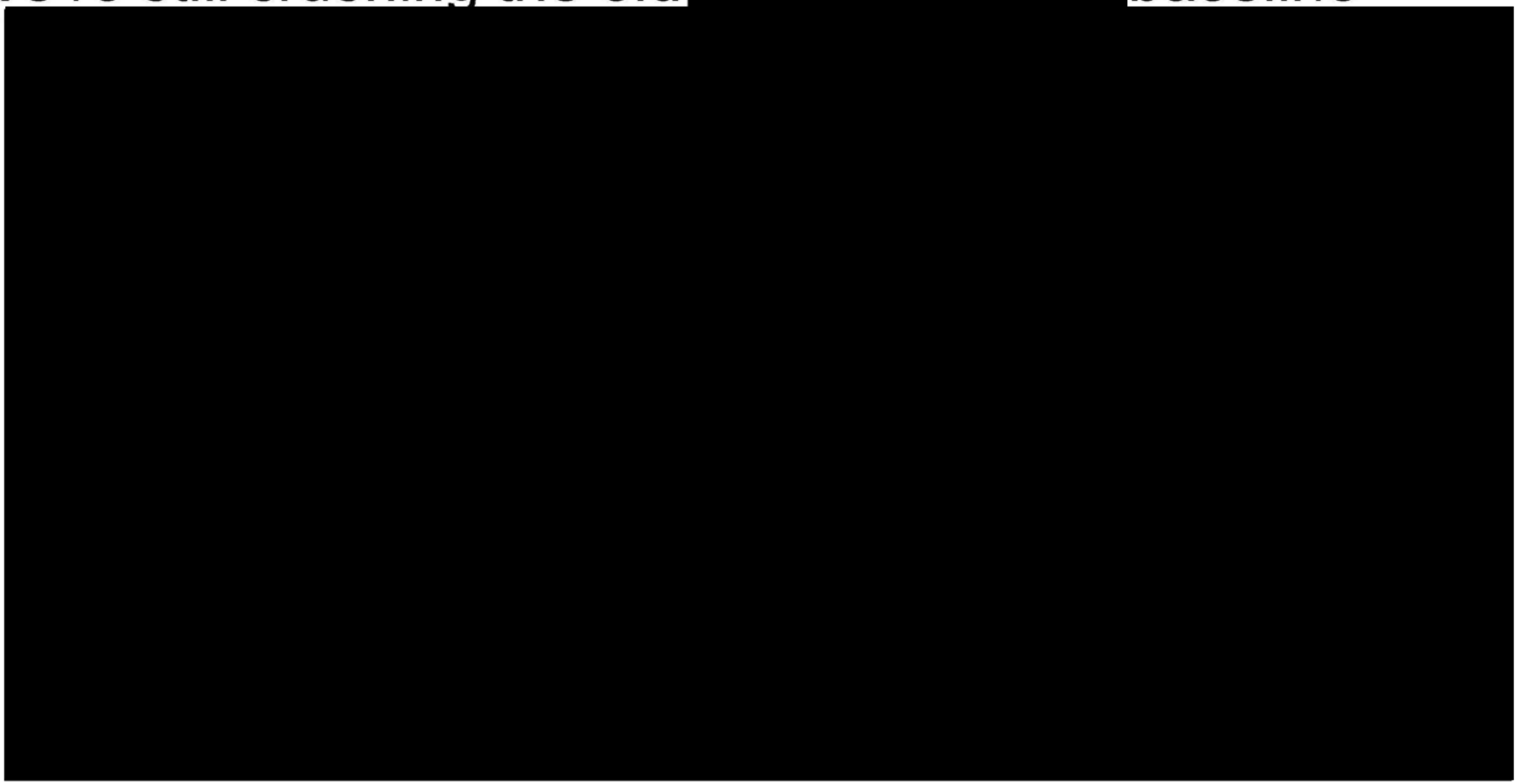
Recall the amazing CM's we were seeing



The mystery of the vanishing CM's



We're still crushing the old [redacted] baseline



What we ended up doing



Q and A

Any questions?