RLHF models & Co-training with Actions Data

Ryan Faulkner

Oct 2024

Motivation

Until now we've largely considered pre-trained model checkpoints when training a policy.

There are LLMs that undergo fine tuning phases after pre-training. In particular instruction tuning and RLHF tuning for model alignment. This leads to models with potentially desirable instruction following and conversational priors.

As we fine tune on actions data, there can be interference with these priors. Can we begin to understand whether we can retain them and what impact this will have on our policy. (DomainX evaluation at the moment)

Definitions

Co-training

- Feed examples from different datasets.
- We have some LLM IT datasets
- Not curated, we have inherited them.
- We have seen that some co-training can keep the model conversational, but what effect does this have on the policy?

RLHF Checkpoints

- Instruction tuned models.
- Are these models are better suited to instruction following in the sense that we define this for a Sima agent?

Confidential Code Excerpt

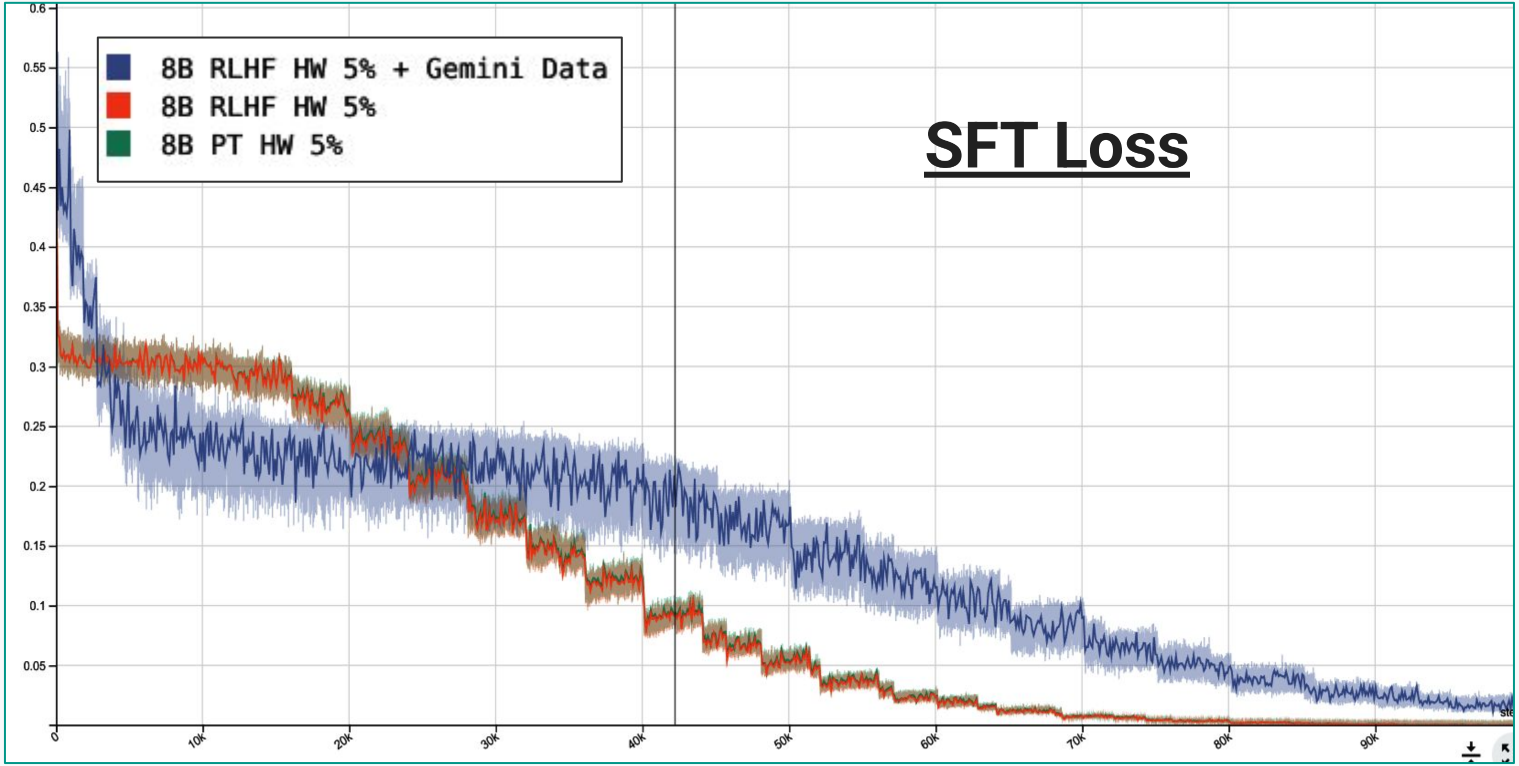
🔗Experiments

- Executed several fine tuning runs against variables in the table below
    - See details @ xxxx
- Common hyper parameters:
    - **TRAIN**: Max steps (depends on dataset + model), LR=1.5e-4, batch size, formatting tokens are included. Splits are 4 to 1 in favour of Domain data.
    - **EVAL**: Temp. = 1.5, Top-k = 100, Top-p = 0.95, Domain v8 Challenge

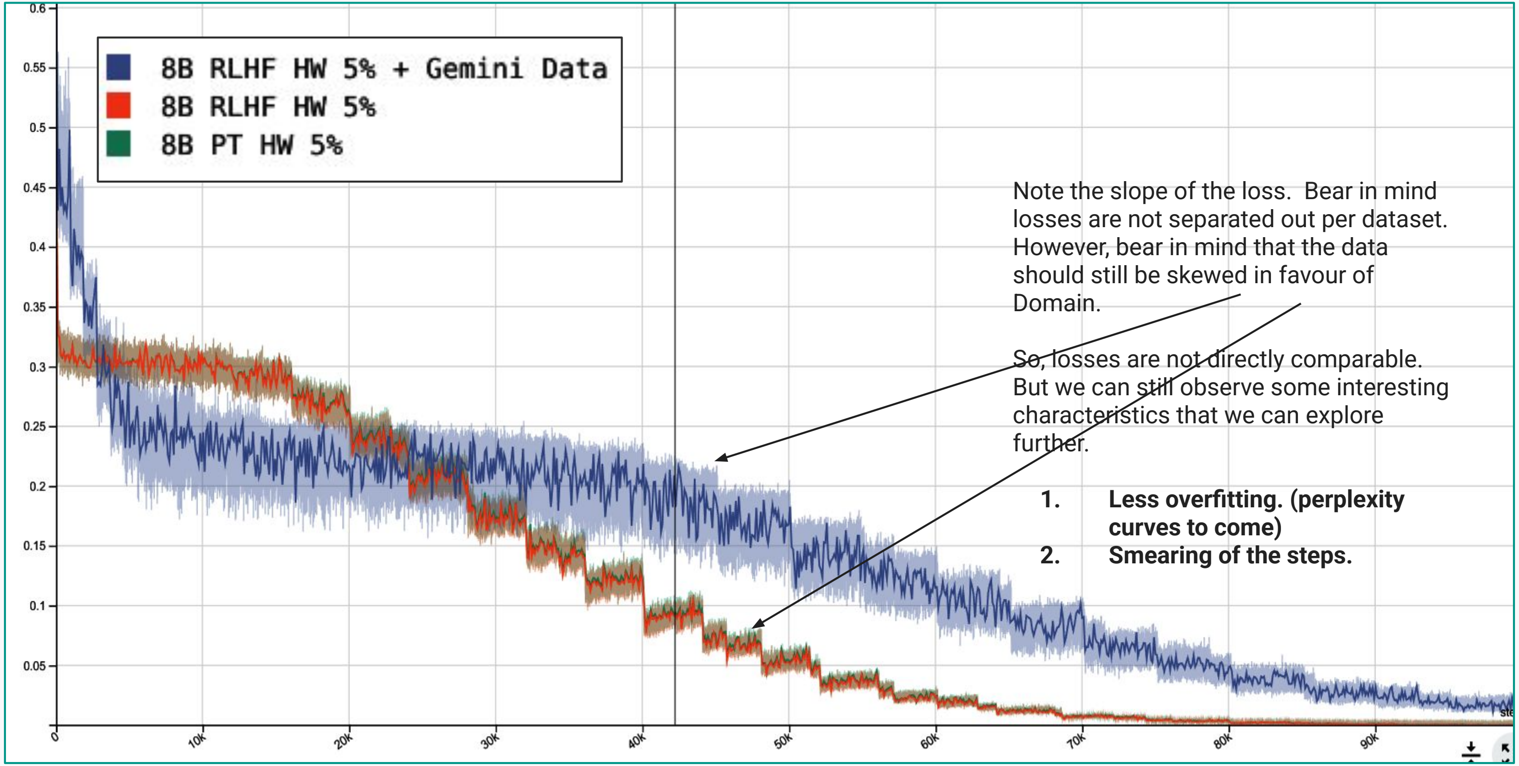
Model Size	Pre Training Phases	Gemini Data	Dataset
{8B, 26B}	{RLHF, PT}	{True, False}*	{DomainX 1%, DomainY 5%}

*** Only True for RLHF and this is actually a non-negative integer value based on Gemini Data Weight**

🔗Domain 5% XS Models - 🔗Evaluate Domain v8



🔗Domain 5% XS Models - 🔗Evaluate Domain v8

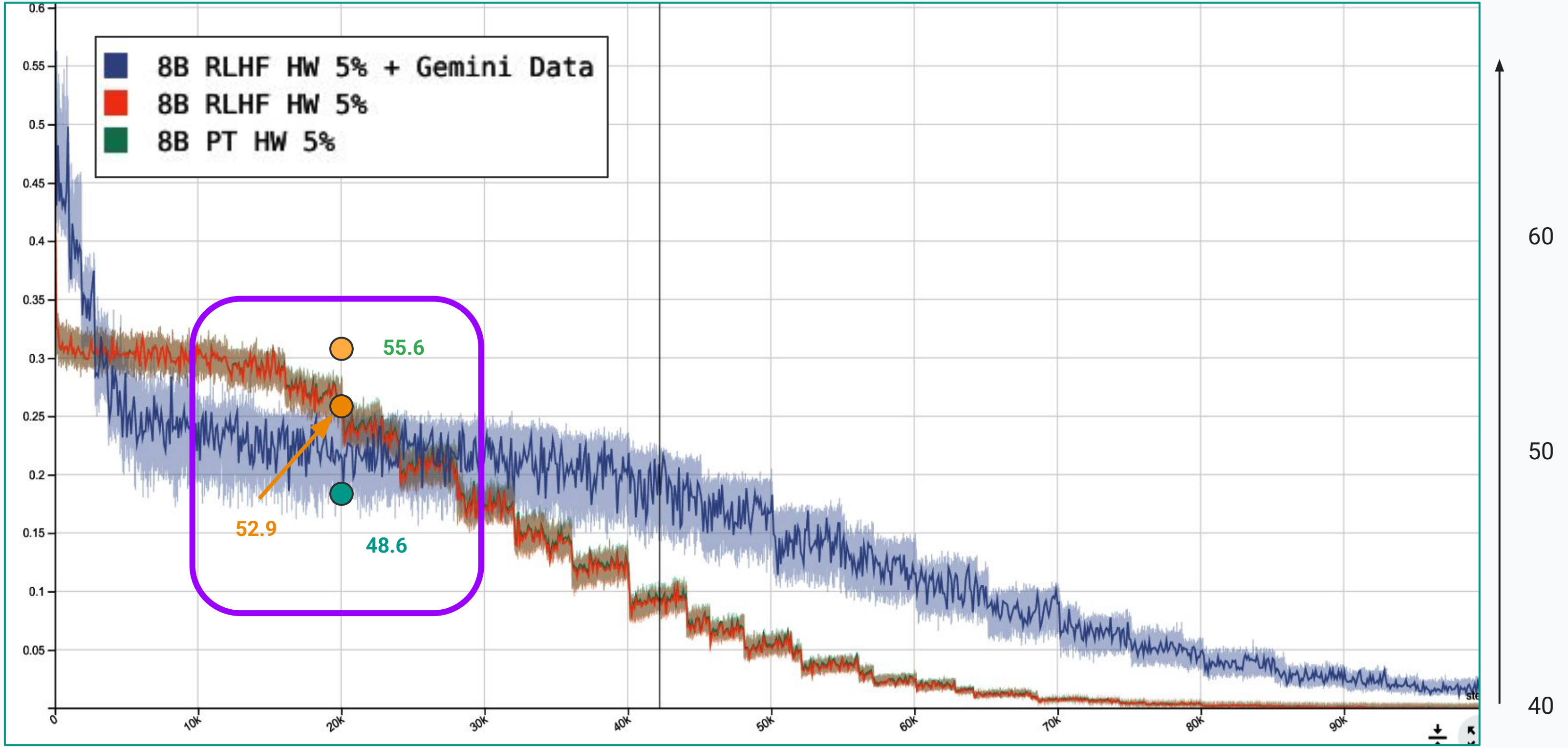


Note the slope of the loss. Bear in mind losses are not separated out per dataset. However, bear in mind that the data should still be skewed in favour of Domain.

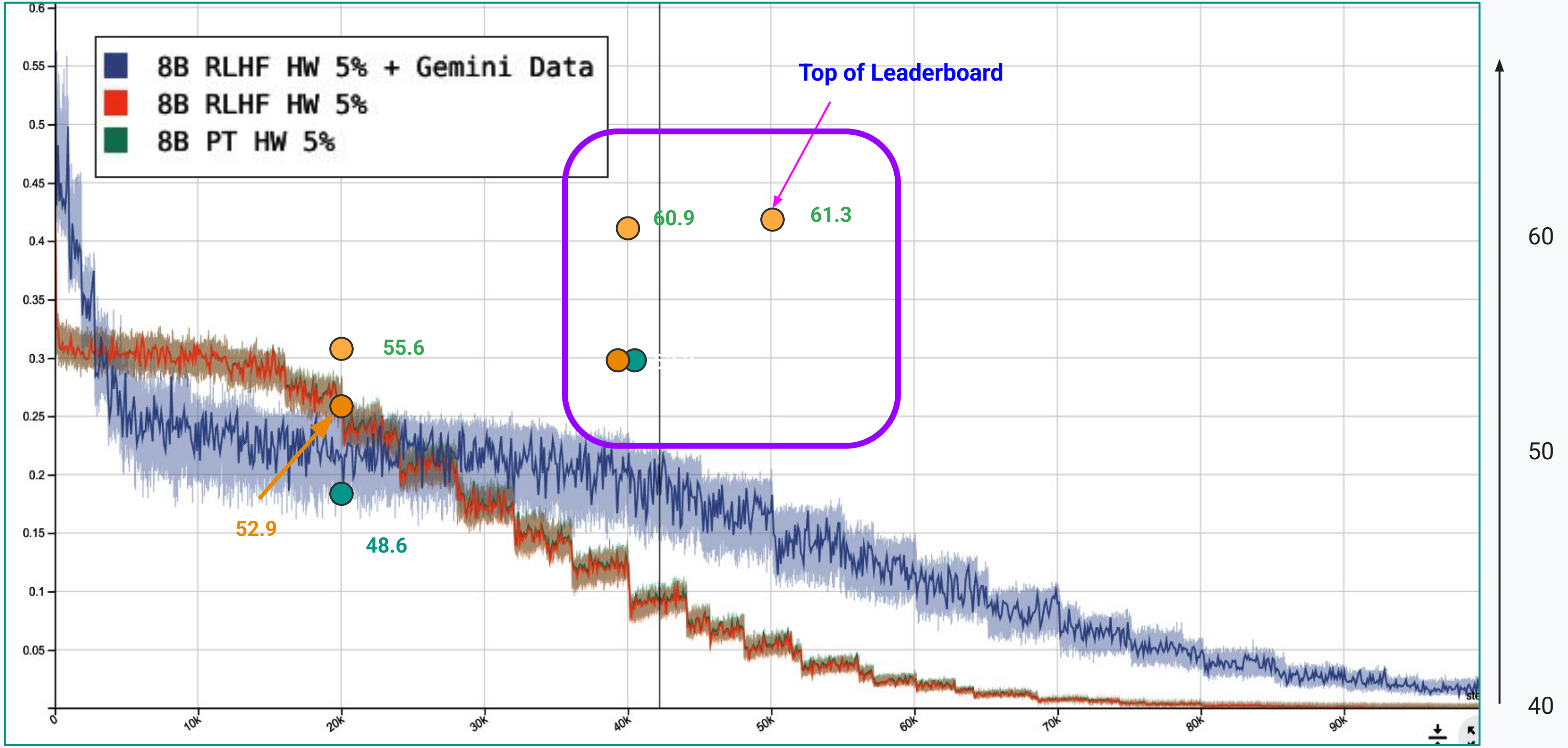
So, losses are not directly comparable. But we can still observe some interesting characteristics that we can explore further.

1. **Less overfitting. (perplexity curves to come)**
2. **Smearing of the steps.**

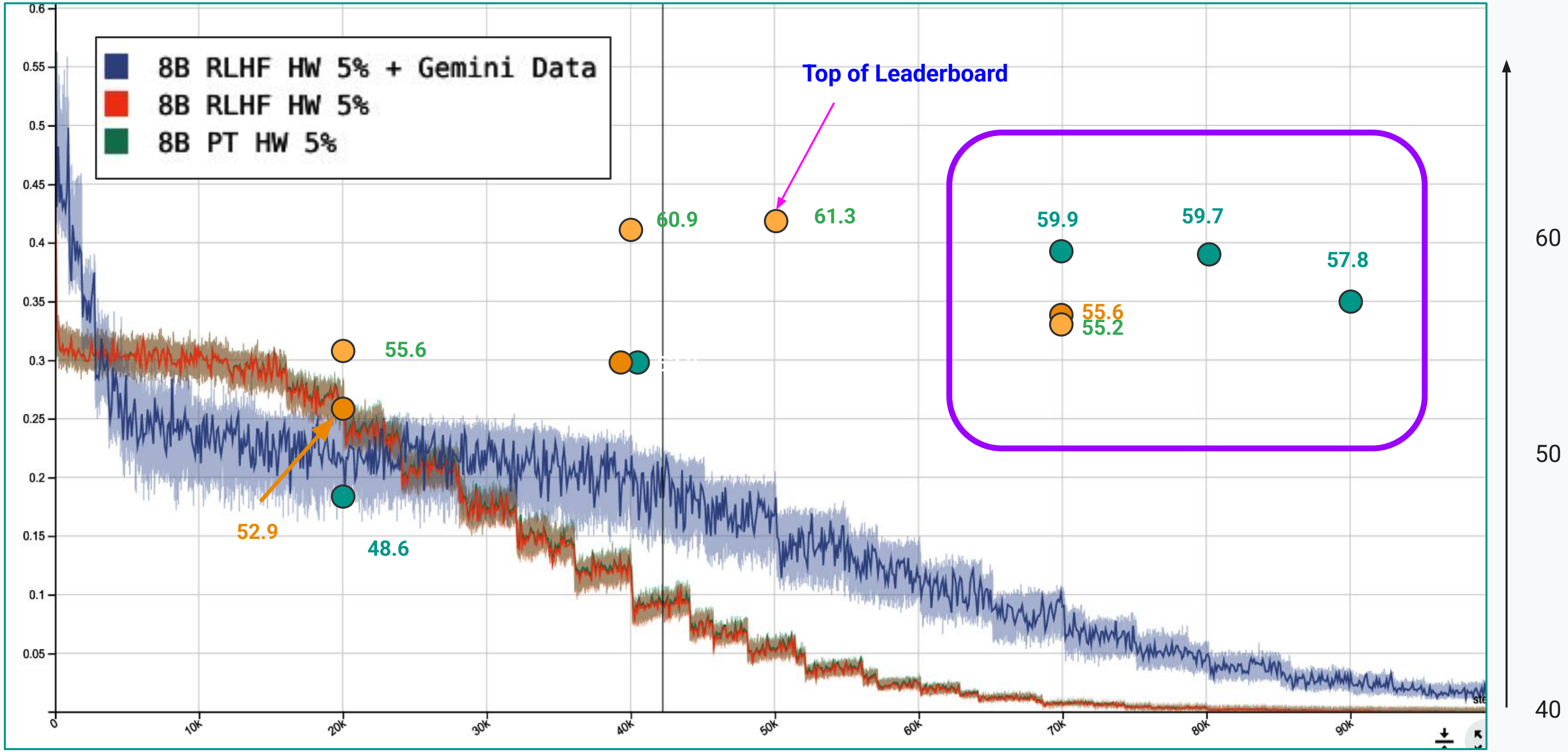
🔗Domain 5% XS Models - 🔗Evaluate Domain v8



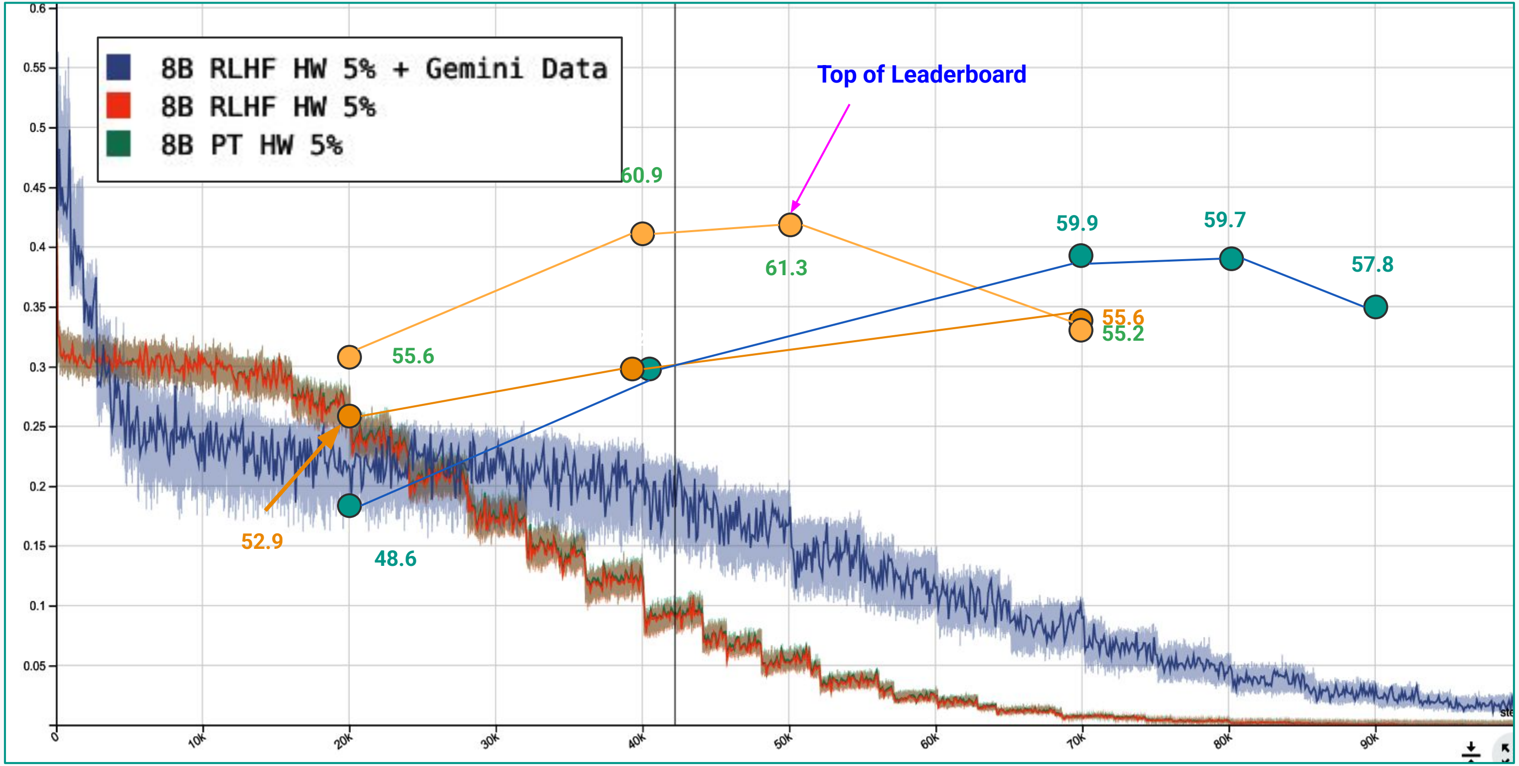
🔗Domain 5% XS Models - 🔗Evaluate Domain v8



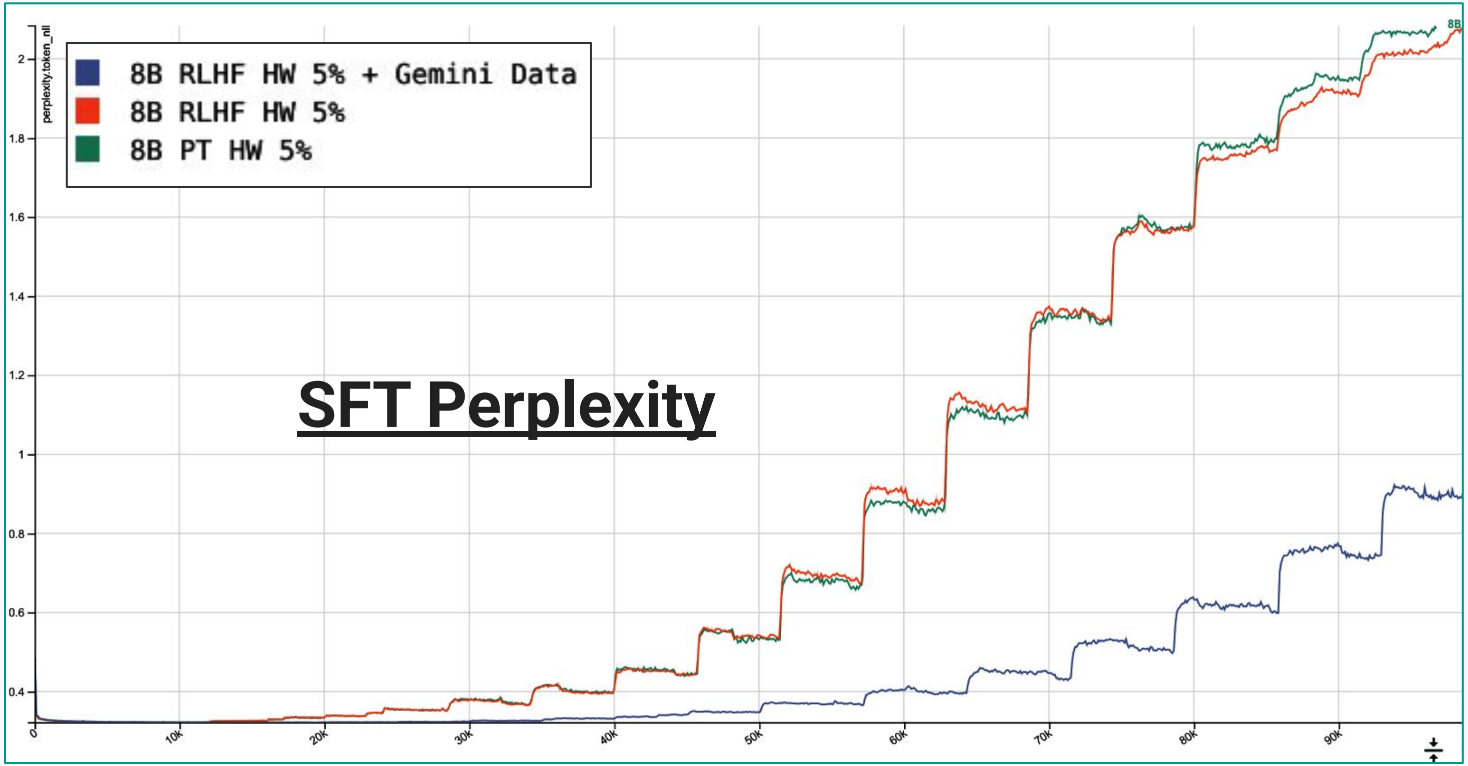
🔗Domain 5% XS Models - 🔗Evaluate Domain v8


Top of Leaderboard

🔗Domain 5% XS Models - 🔗Evaluate Domain v8

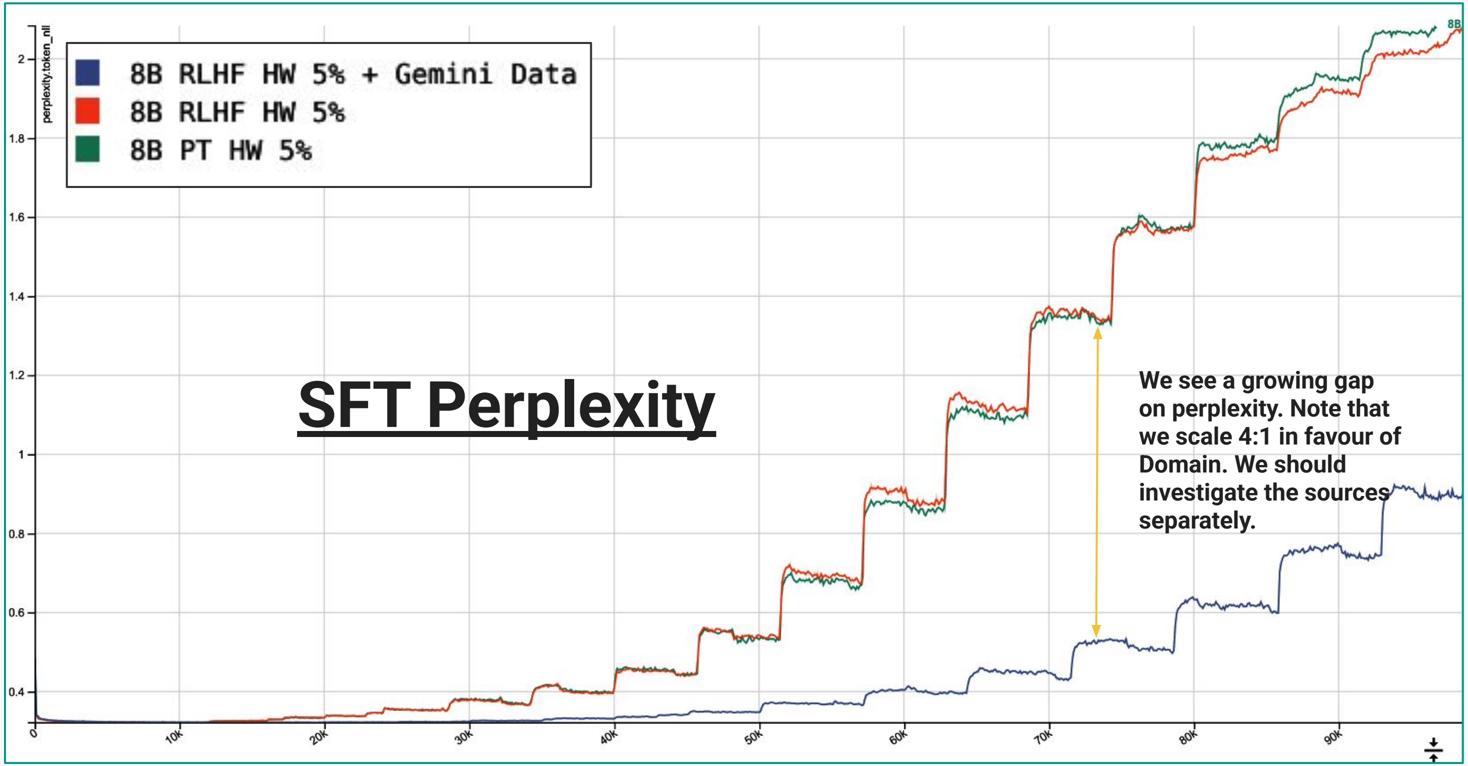


Domain 5% XS Models - 🔗Evaluate Domain v8



SFT Perplexity

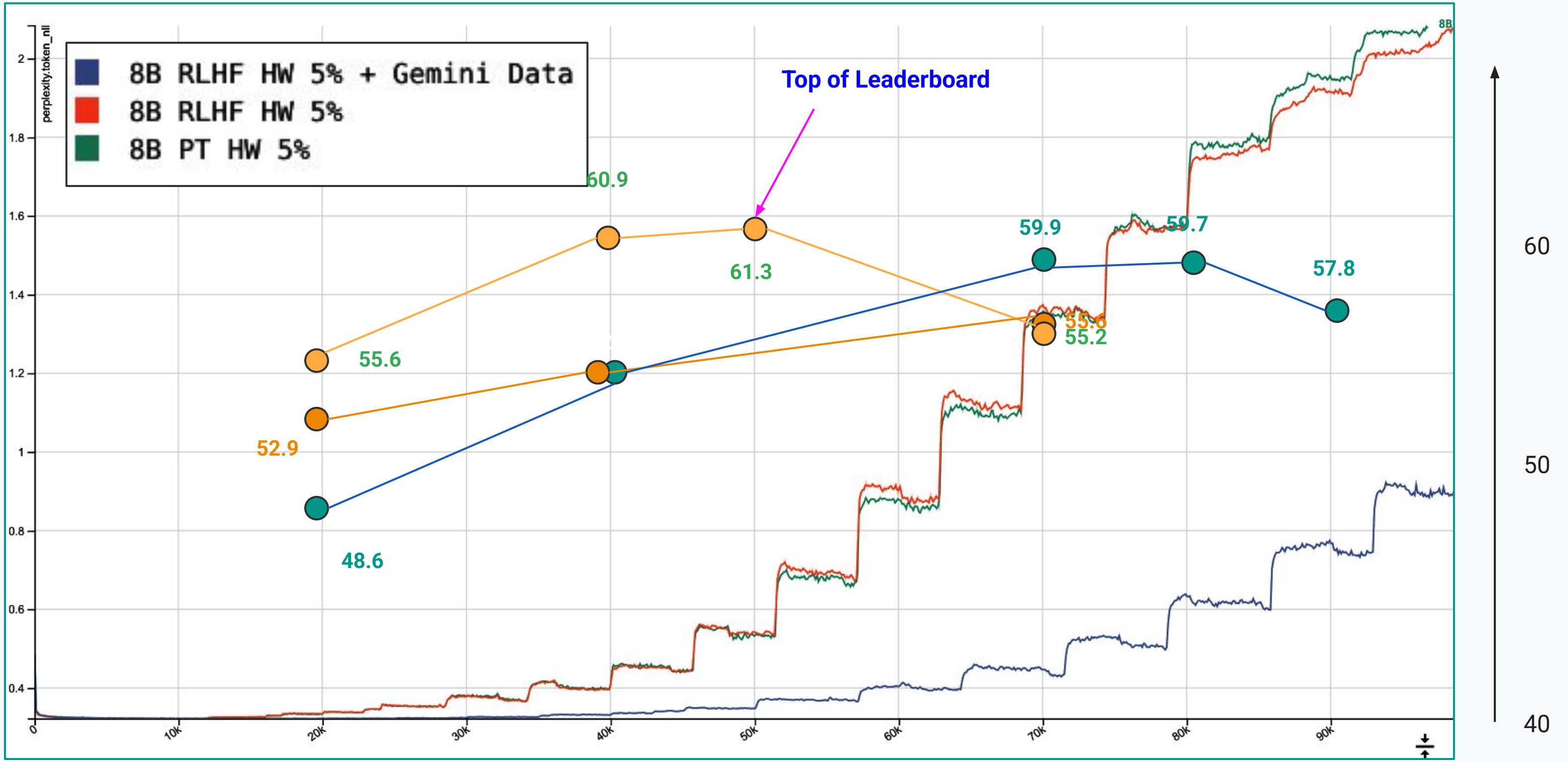
🔗Domain 5% XS Models - 🔗Evaluate Domain v8



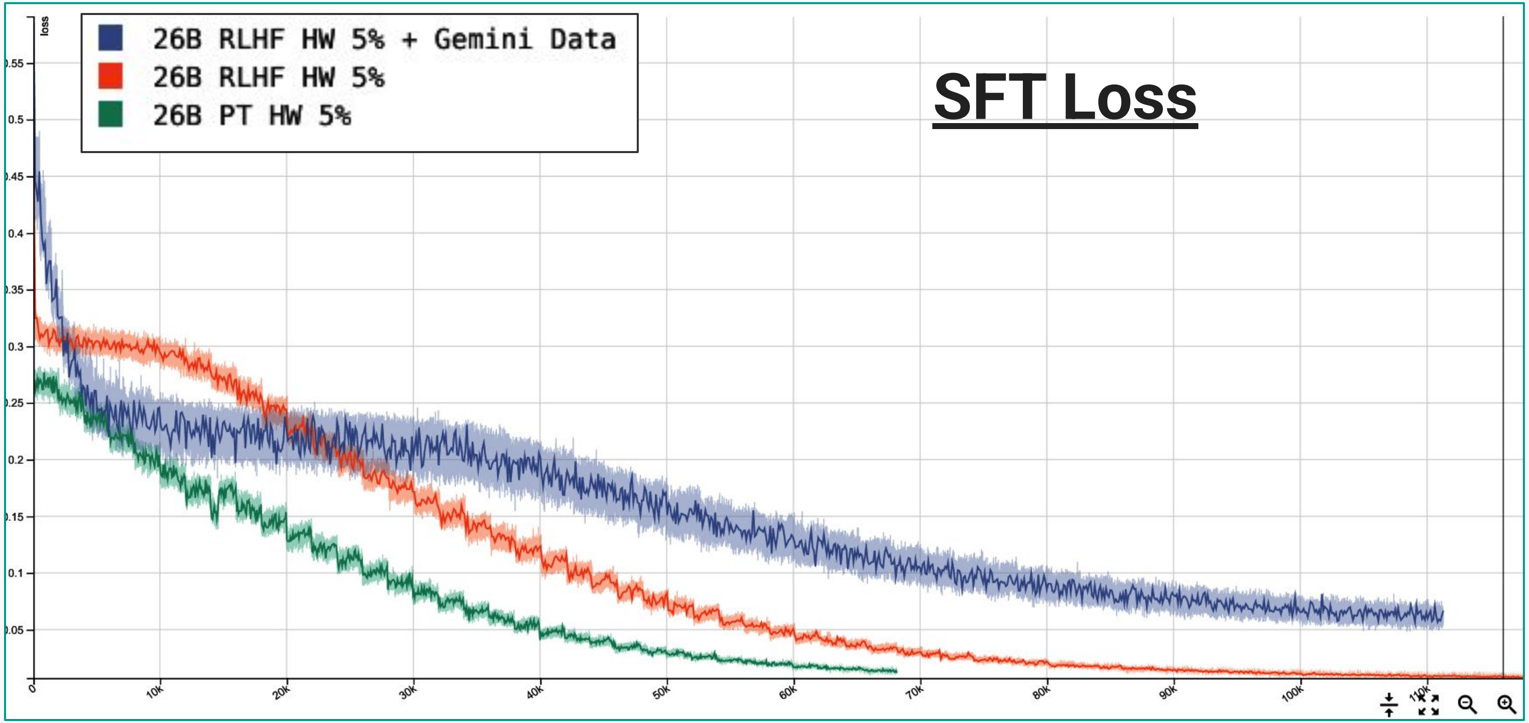
SFT Perplexity

We see a growing gap on perplexity. Note that we scale 4:1 in favour of Domain. We should investigate the sources separately.

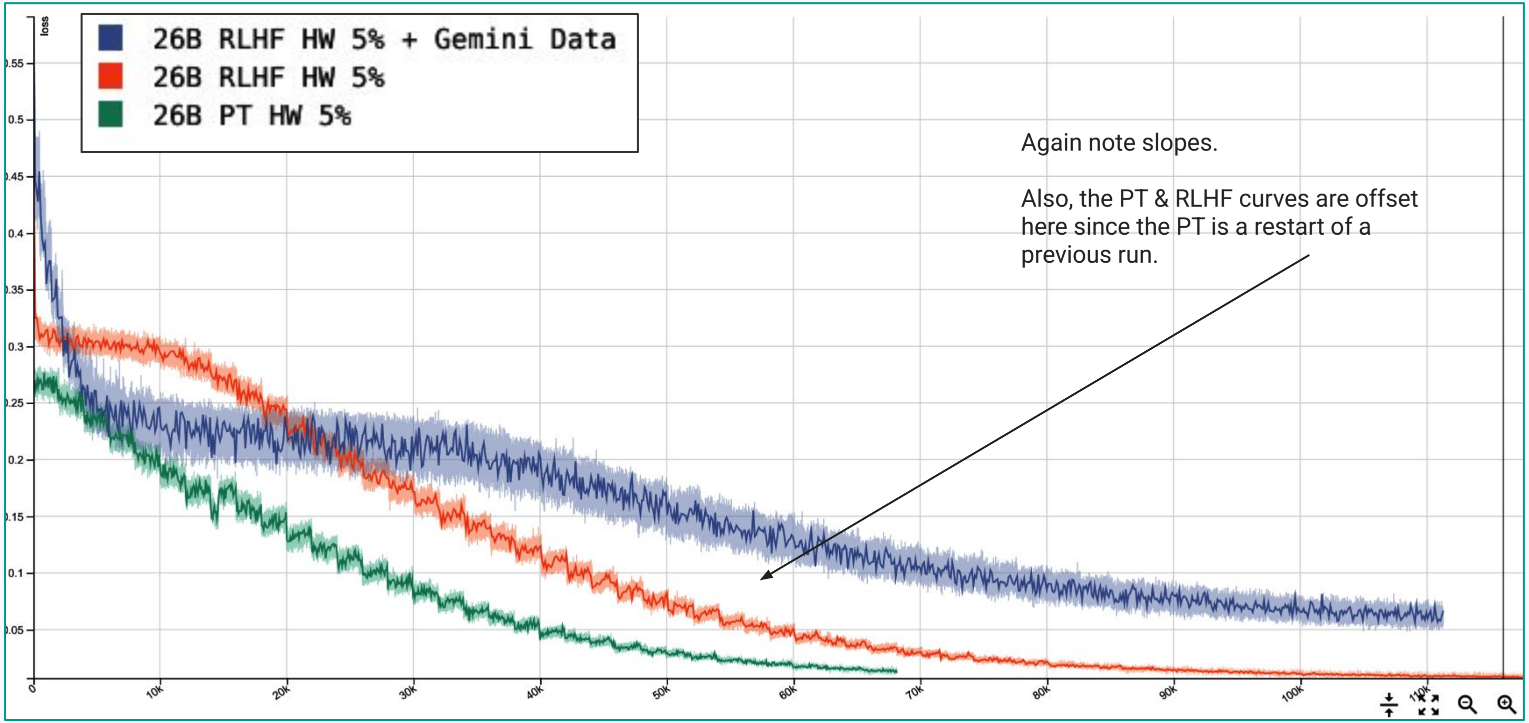
🔗Domain 5% XS Models - 🔗Evaluate Domain v8



🔗Domain 5% S Models - 🔗Evaluate Domain v8



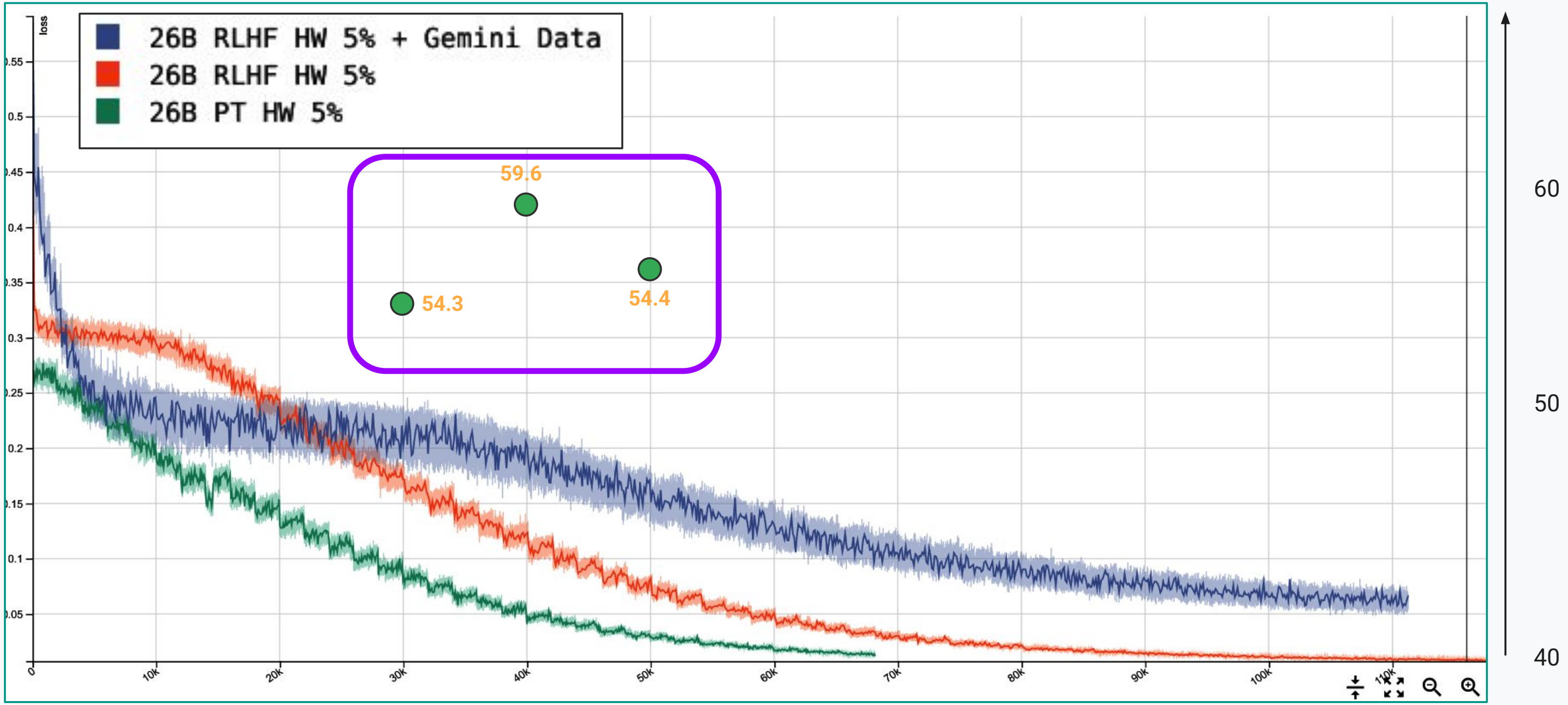
🔗Domain 5% S Models - 🔗Evaluate Domain v8



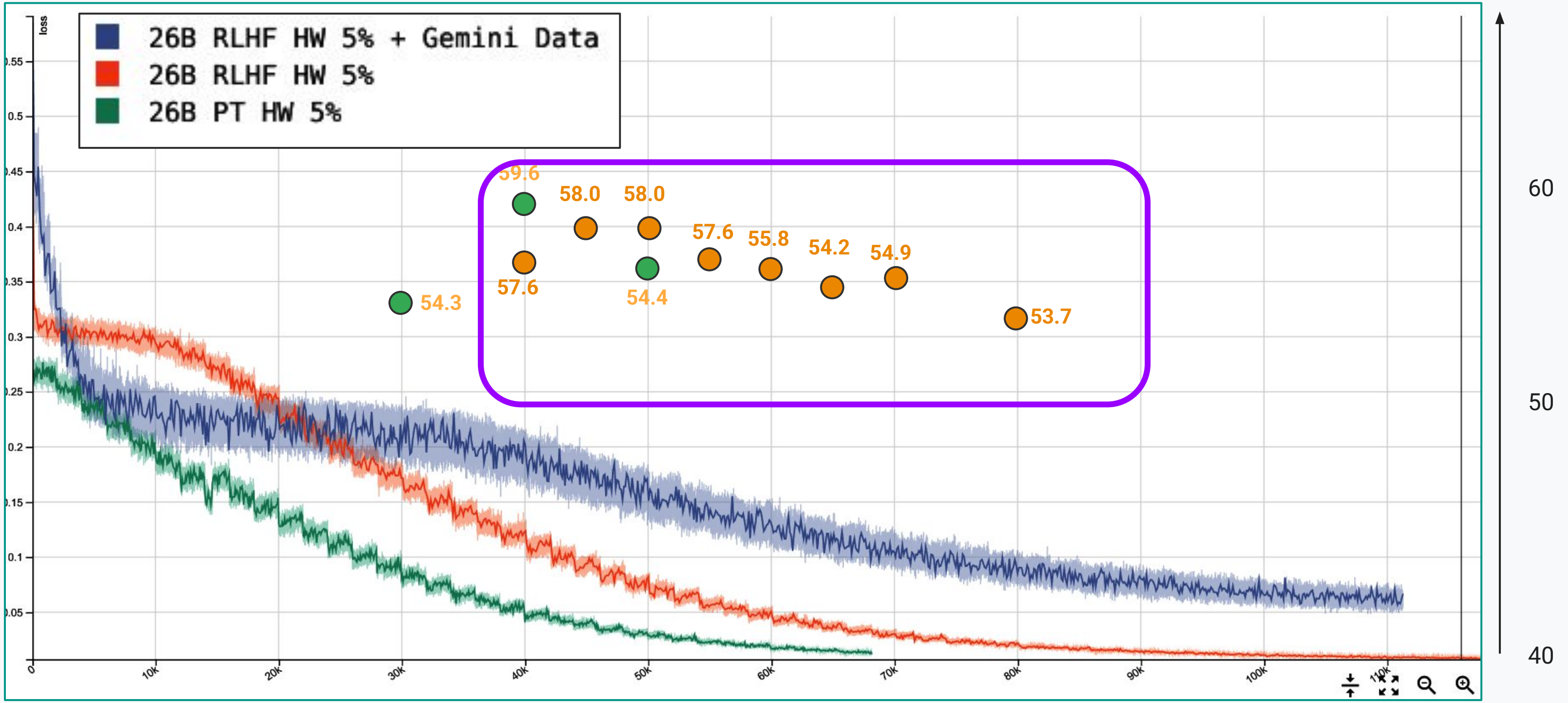
Again note slopes.

Also, the PT & RLHF curves are offset here since the PT is a restart of a previous run.

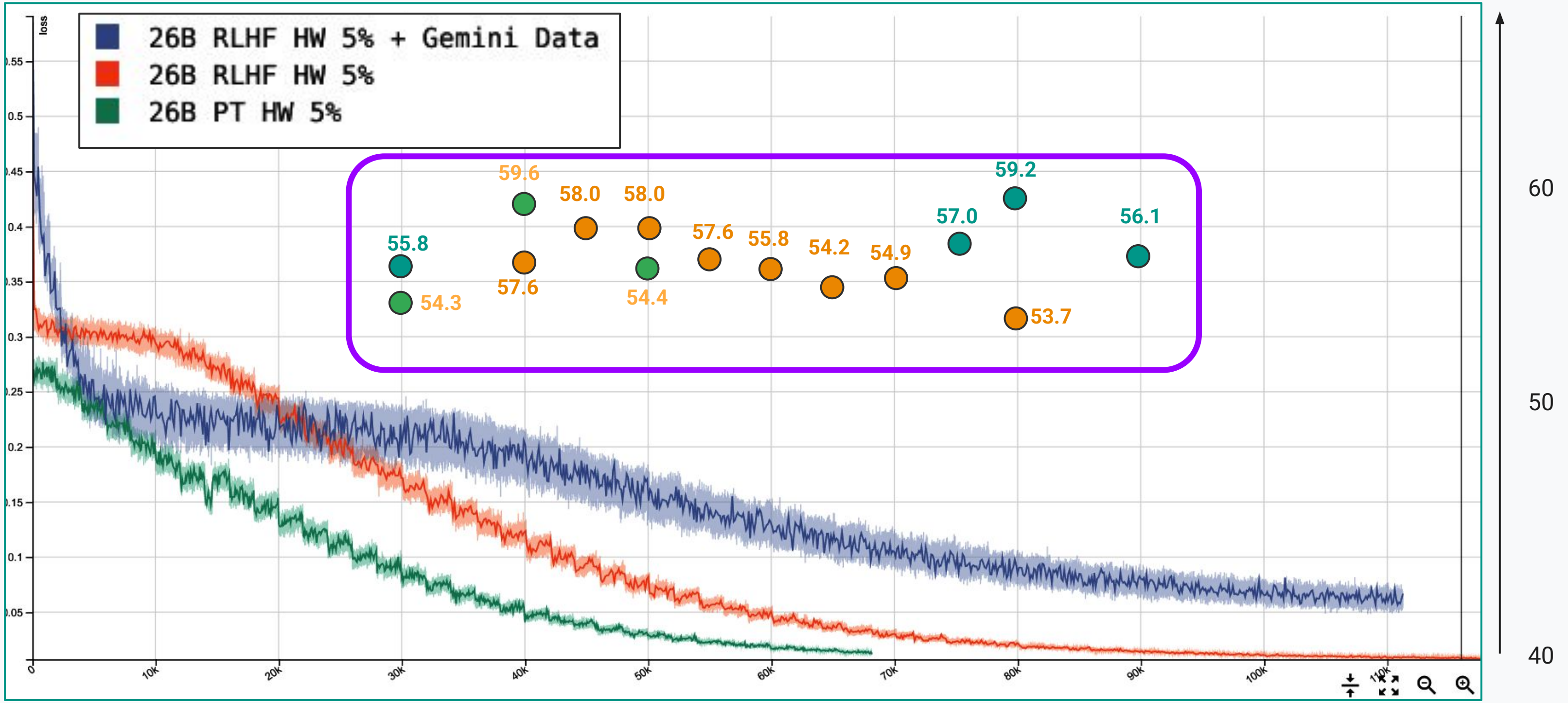
🔗Domain 5% S Models - 🔗Evaluate Domain v8



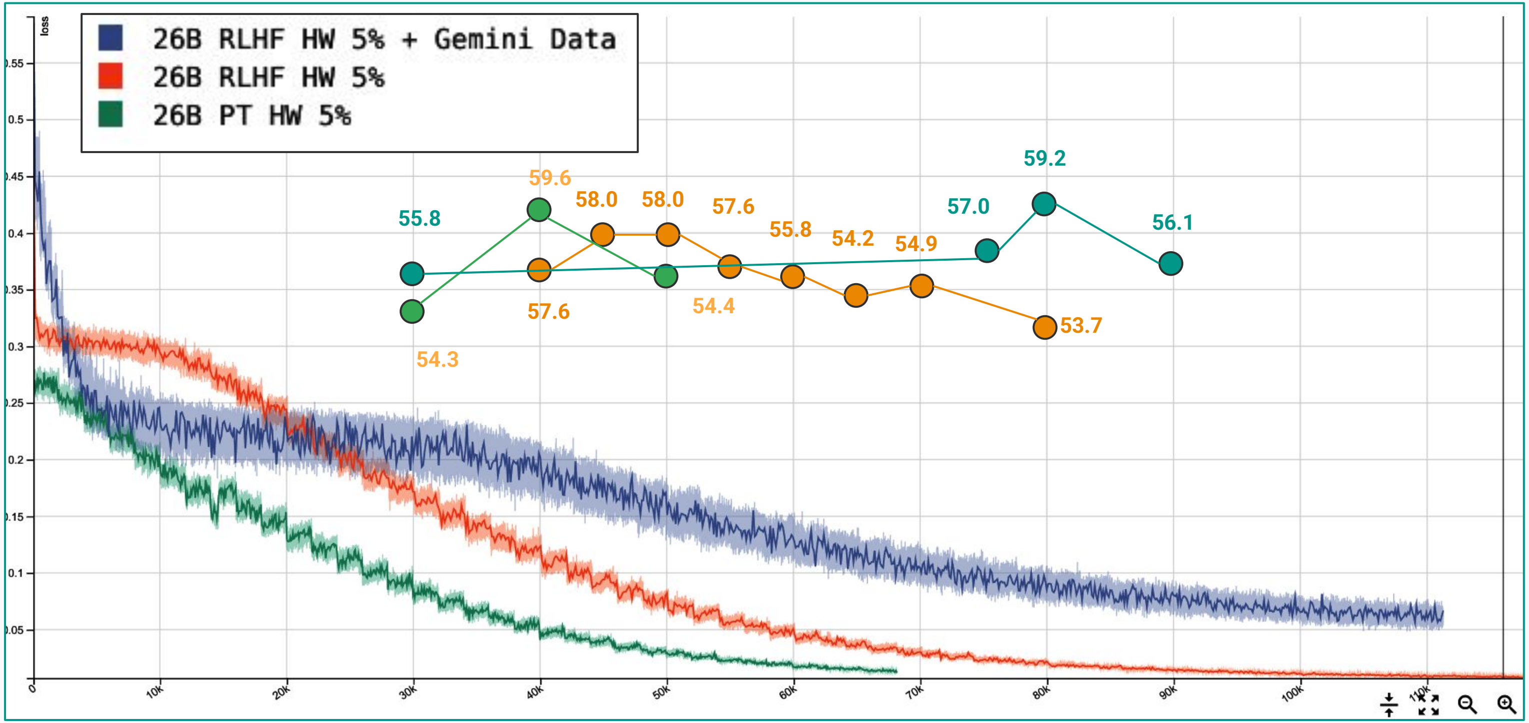
🔗Domain 5% S Models - 🔗Evaluate Domain v8



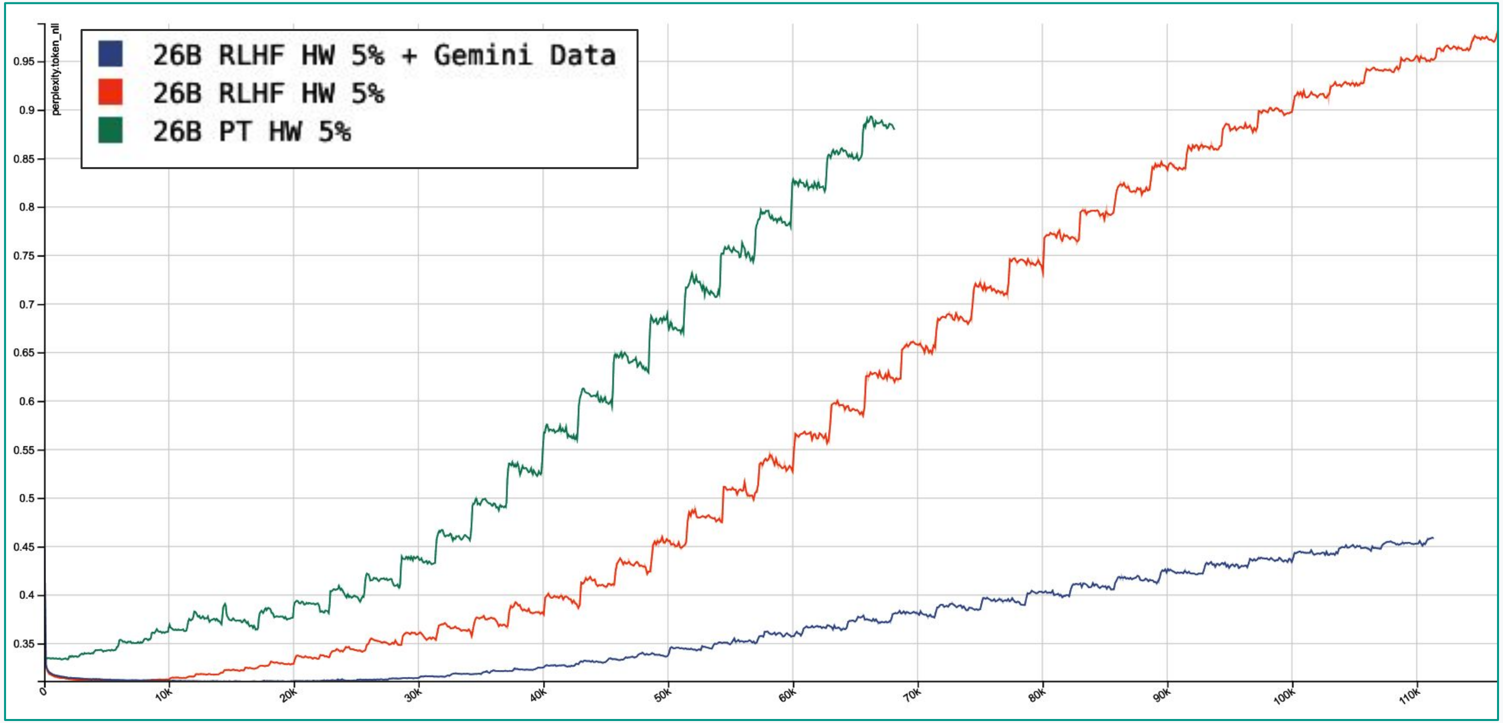
🔗Domain 5% S Models - 🔗Evaluate Domain v8



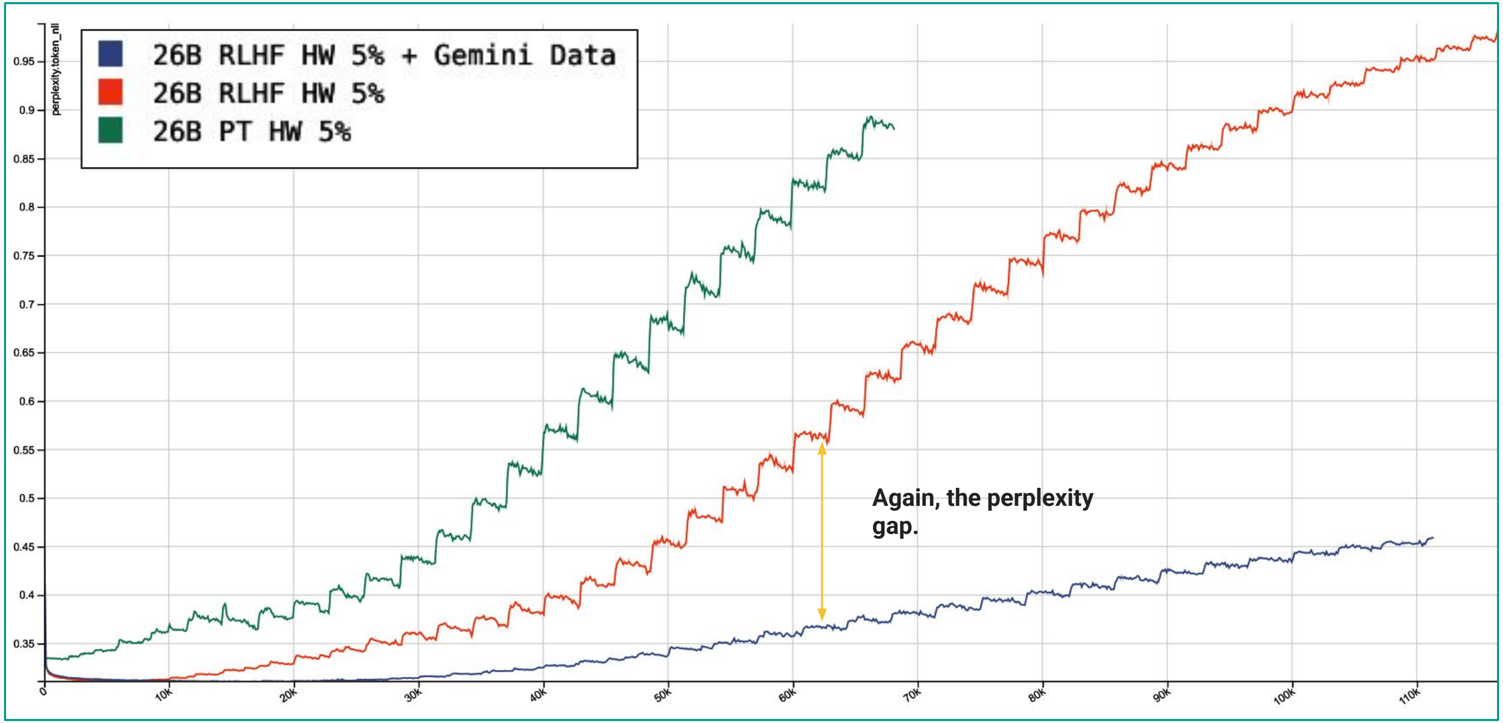
🔗Domain 5% S Models - 🔗Evaluate Domain v8



🔗Domain 5% S Models - 🔗Evaluate Domain v8

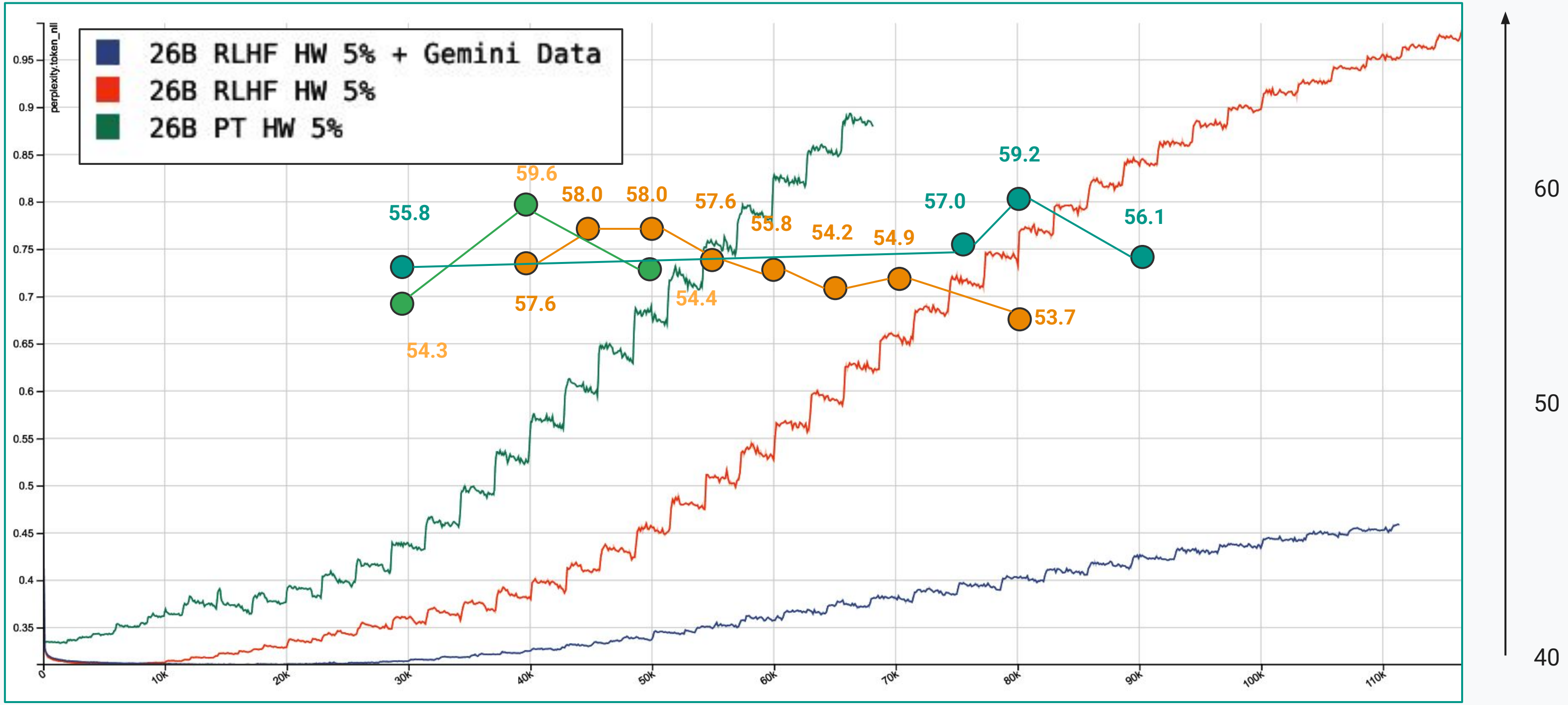


🔗Domain 5% S Models - 🔗Evaluate Domain v8

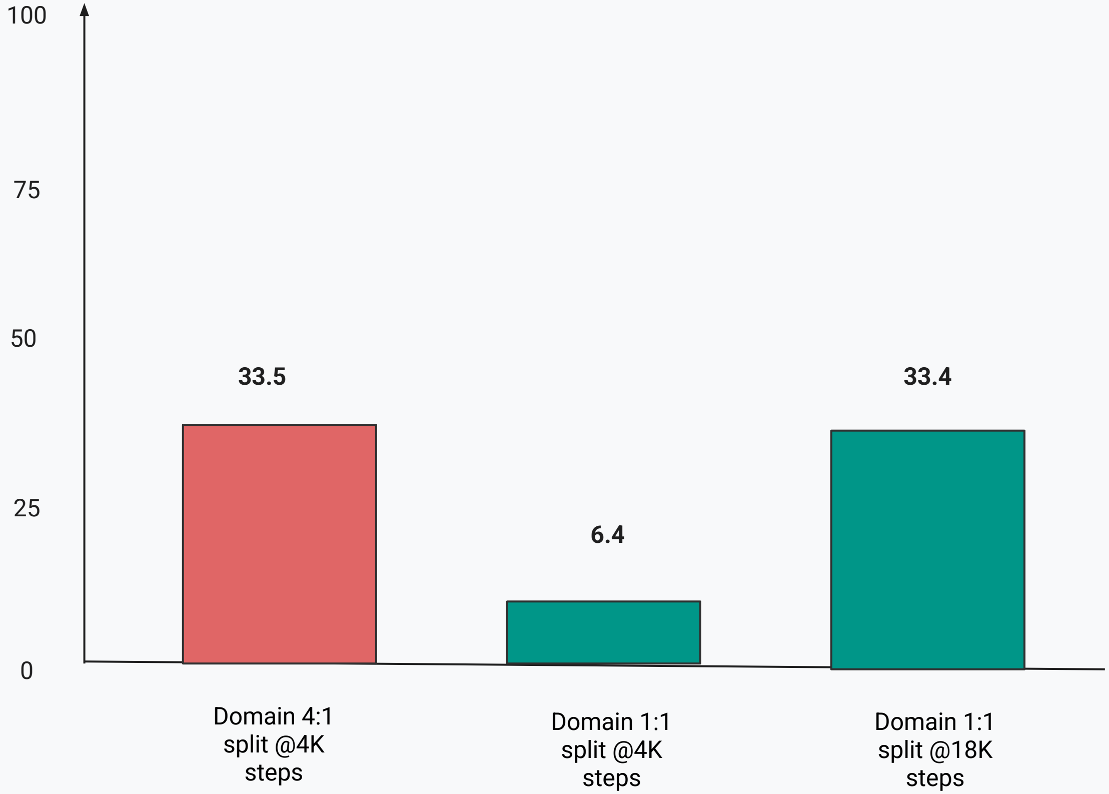


Again, the perplexity gap.

🔗Domain 5% S Models - 🔗Evaluate Domain v8



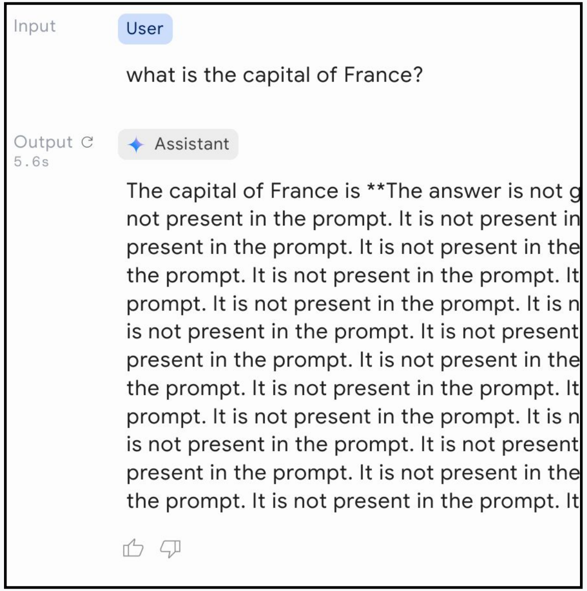
Comparing Ratio of LLM Data on Domain 1%



- Does a higher ratio LLM data in the mix affect convergence?
- We have already seen that convergence is slowing in the presence of LLM IT datasets on a 4:1 weight split in favour of Domain
- Eventually we can match performance with a higher ratio but convergence takes longer.

Language Request @ 4K steps

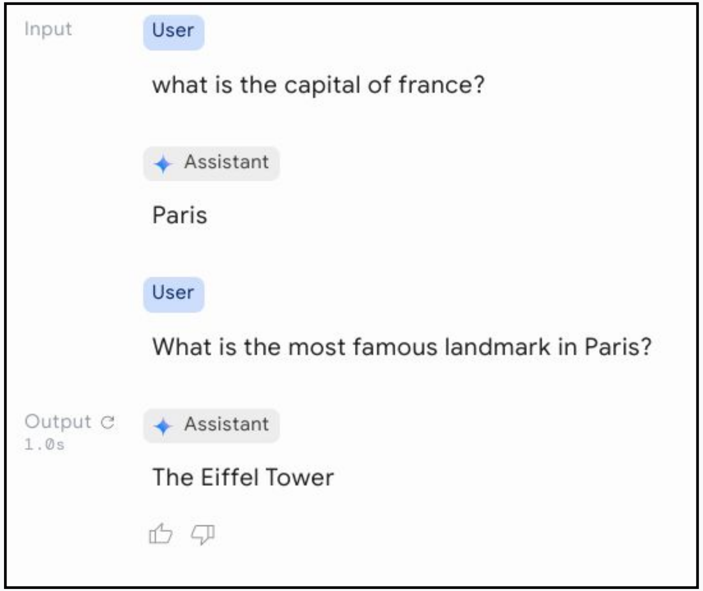
RHLF SFT Domain



what is the capital of France?

The capital of France is **The answer is not g
not present in the prompt. It is not present in
present in the prompt. It is not present in the
the prompt. It is not present in the prompt. It
prompt. It is not present in the prompt. It is n
is not present in the prompt. It is not present
present in the prompt. It is not present in the
the prompt. It is not present in the prompt. It
prompt. It is not present in the prompt. It is n
is not present in the prompt. It is not present
present in the prompt. It is not present in the
the prompt. It is not present in the prompt. It

RHLF SFT Domain+LLM 1:1



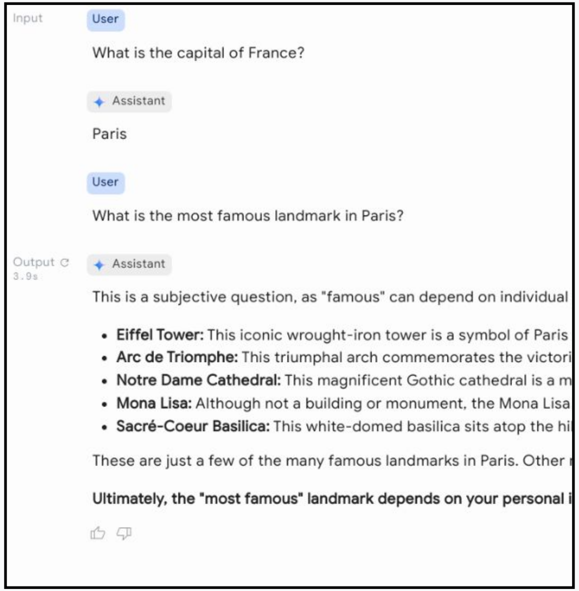
what is the capital of france?

Paris

What is the most famous landmark in Paris?

The Eiffel Tower

RHLF SFT Domain+LLM 4:1



What is the capital of France?

Paris

What is the most famous landmark in Paris?

This is a subjective question, as "famous" can depend on individual

- **Eiffel Tower:** This iconic wrought-iron tower is a symbol of Paris
- **Arc de Triomphe:** This triumphal arch commemorates the victori
- **Notre Dame Cathedral:** This magnificent Gothic cathedral is a m
- **Mona Lisa:** Although not a building or monument, the Mona Lisa
- **Sacré-Coeur Basilica:** This white-domed basilica sits atop the hi

These are just a few of the many famous landmarks in Paris. Other

Ultimately, the "most famous" landmark depends on your personal

Prior Work: Summary of conversation checks

- PT, initial: no (nonsensical)
- PT, fine tuning, no co-training: no (nonsensical)
- Capable model but serving without ctrl tokens / formatting: no (prompt completion)
- RLHF, initial: yes
- RLHF, fine tuning, co-training: yes
- PT, fine tuning, co-training: yes (!)
    - I.e.: the co-training data induces conversational abilities
- Does the RLHF model forget if not co-trained?
    - Yes, starting with complex queries (“tell me history of US”) then simpler (“1+1”)

Takeaways

- 👍 Co-training doesn't seem to have a significant negative impact on Domain performance.
- 👍Co-training helps retain language capabilities in the model. Co-training might even improve things a little for the policy.
    - 🔬If we think of this process as a regulariser, we might use better aligned datasets for policy optimisation (e.g. GQA etc.). We may actually be able to improve generalisation of the policy.
    - 🔬Run conversational metrics or challenges
    - 🔬Does conditioning on language the policy (e.g. rationales)
- 🔬Model convergence is affected by relative weight of co-training data.
- 🔬Experiments were done with formatting tokens. We should measure the effect of them in a controlled experiment.