Small Transformers Don't Need LayerNorm at Inference Time

Luca Baroni*, Galvin Khara*, Joachim Schaeffer*, Marat Subkhankulov* and Stefan Heimersheim

PAPER CODE MODELS

ICLR

Motivation

- LayerNorm is a non-linearity that makes **mechanistic interpretability harder**
- LayerNorm may not be essential for inference

Key results

- LayerNorm is removed from GPT-2 Small to XL
- Performance of LN-free models closely matches the original
- LN-free models have interesting properties wrt mechanistic interpretability

LayerNorm definition

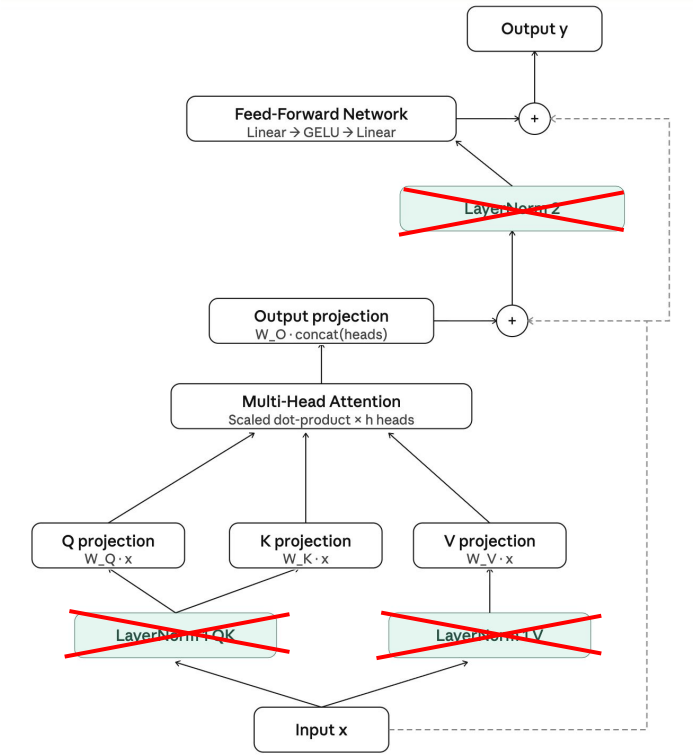
$$\mathbf{LN}(\mathbf{x}) = \frac{\mathbf{x} - \mu}{\sigma} \odot \boldsymbol{\gamma} + \boldsymbol{\beta}, \quad \mu = \frac{1}{H}\sum_{h=1}^{H} x_h, \quad \sigma = \sqrt{\frac{1}{H}\sum_{h=1}^{H}(x_h - \mu)^2 + \epsilon},$$

LayerNorm removal

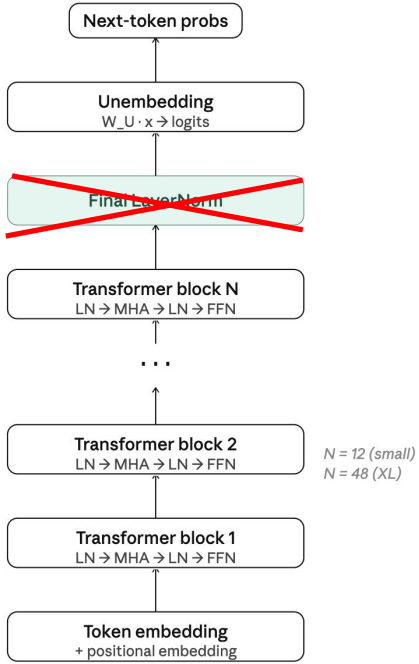
- Linearize LN by freezing the standard deviation
- While fine-tuning on original training data (OWT)
- Gradual schedule to maintain CE loss

$$\text{FakeLN}(\mathbf{x}) = \frac{\mathbf{x} - \mu}{\overline{\sigma}_{\text{avg}}} \odot \boldsymbol{\gamma} + \boldsymbol{\beta}, \quad \sigma_{b,s} = \sqrt{\frac{1}{H}\sum_{h=1}^{H}(x_{b,s,h} - \mu_{b,s})^2 + \epsilon}, \quad \sigma_{\text{avg}} = \frac{1}{BS}\sum_{b=1}^{B}\sum_{s=1}^{S}\sigma_{b,s},$$

LayerNorm removal

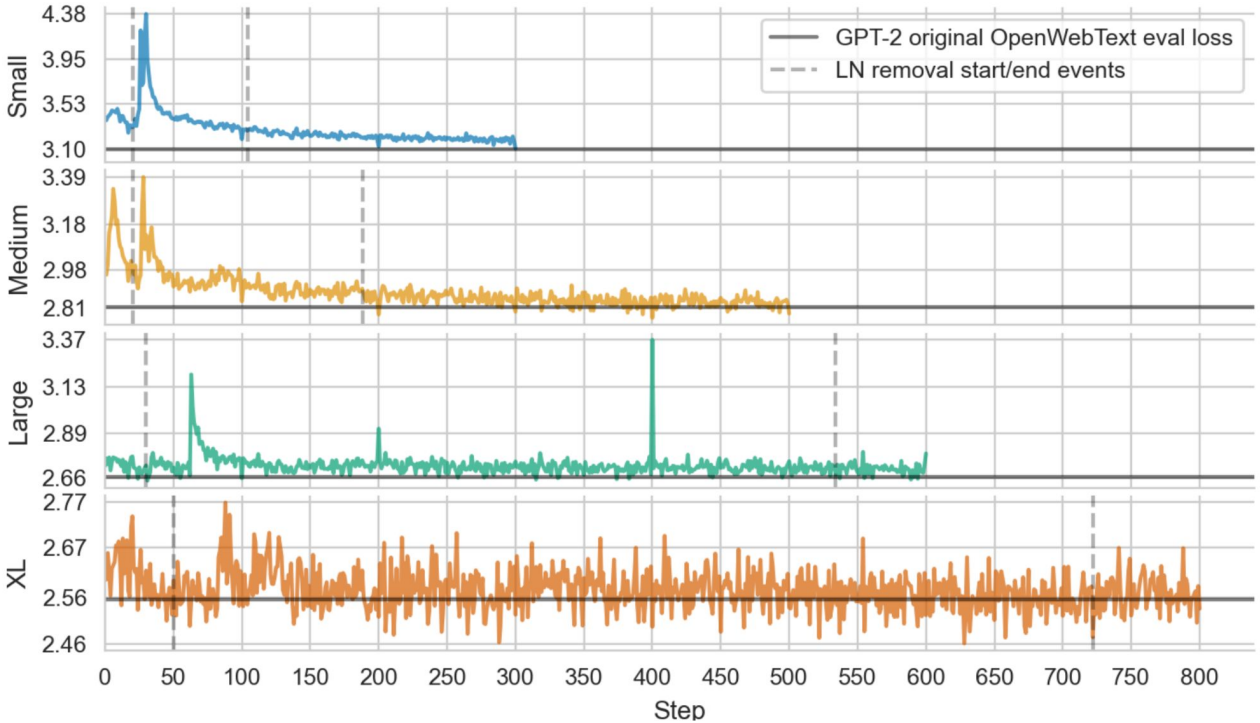


GPT-2 transformer block



GPT-2 transformer

LayerNorm removal



LayerNorm removal

- EMA of $\overline{\sigma}_{\text{avg}}$
- Auxiliary loss to push standard deviations closer to the average
- Split QK and V LayerNorm removal

$$\mathcal{L}_{\text{aux}} = \lambda \cdot \mathbb{E}_{b,s}\left[(\sigma_{b,s} - \hat{\sigma})^2\right], \qquad \hat{\sigma} = \frac{1}{|\mathcal{M}|} \sum_{(b,s)\in\mathcal{M}} \sigma_{b,s},$$

Evaluation on CE loss

Model	FT steps	OWT (val)	The Pile	The Pile-filtered
GPT-2 Small original	0	3.1006	**2.8450**	**2.7899**
GPT-2 Small vanilla	300	**3.0126**	2.8511	2.8112
GPT-2 Small LN-free	300	3.0797 [+0.0671]	2.8852 [+0.0402]	2.8757 [+0.0858]
GPT-2 Medium original	0	2.8145	**2.5163**	**2.5390**
GPT-2 Medium vanilla	500	**2.7390**	2.5752	2.5724
GPT-2 Medium LN-free	500	2.7642 [+0.0252]	2.6579 [+0.1416]	2.6352 [+0.0962]
GPT-2 Large original	0	2.6623	**2.5320**	**2.4347**
GPT-2 Large vanilla	600	**2.6240**	2.6233	2.5074
GPT-2 Large LN-free	600	2.6384 [+0.0144]	2.7504 [+0.2184]	2.5159 [+0.0812]
GPT-2 XL original	0	2.5567	**2.4436** [3]	**2.3739**
GPT-2 XL Vanilla	800	**2.4799**	2.4673	2.3821
GPT-2 XL LN-free	800	2.5052 [+0.0253]	130.2197 [4]	2.3992 [+0.0253]

[3] GPT-2 XL original: Median: 1.0103, 95th perc: [0.0005, 10.6193], 99.9th perc: [≈0.0000, 43.0064]
[4] GPT-2 XL LN-free: Median: 1.0937, 95th perc: [0.0004, 10.7548], 99.9th perc: [≈0.0000, 48.6459]

Evaluation on standard benchmarks

Task	GPT-2 XL original	GPT-2 XL vanilla FT	GPT-2 XL LN-free FT
BoolQ	61.8	**62.2**	61.9
HellaSwag	**50.9**	49.8	48.8
PIQA	**70.5**	70.5	69.9
WinoGrande	**58.3**	57.5	56.1

Table 6: Accuracy of GPT-2 XL model variants on BoolQ, HellaSwag, PIQA, and WinoGrande.

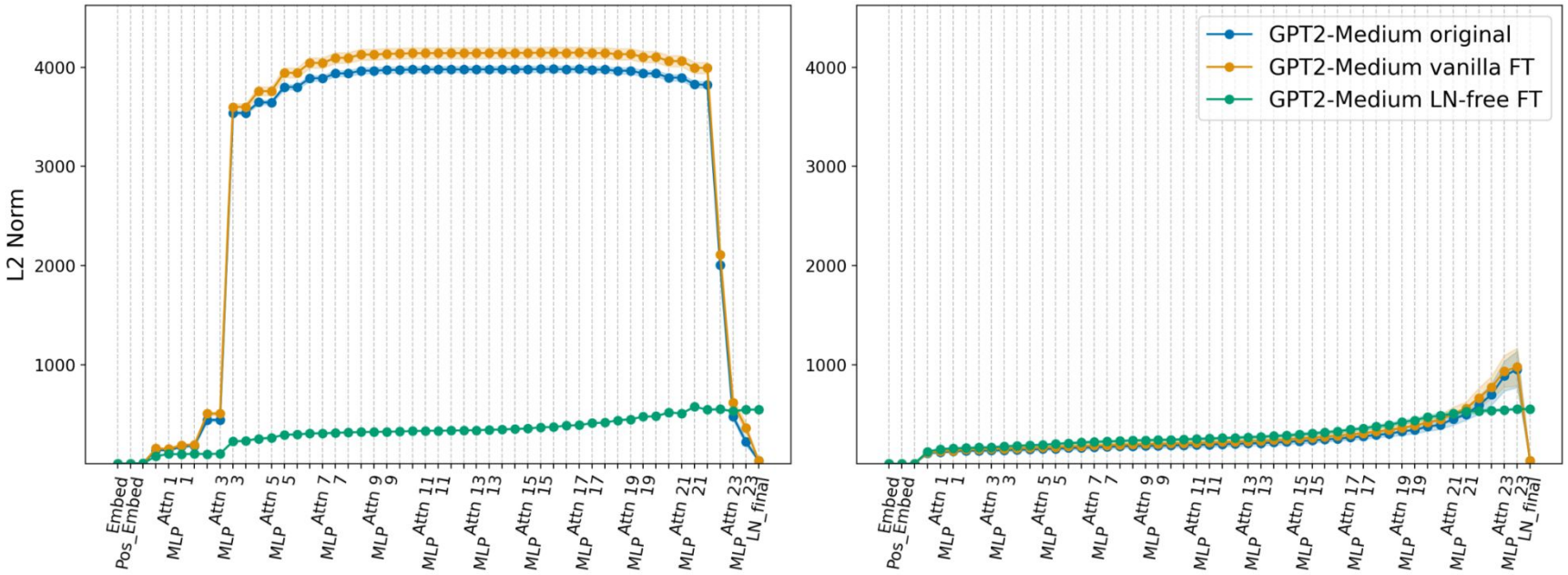
Task	GPT-2 Large original	GPT-2 Large vanilla FT	GPT-2 Large LN-free FT
BoolQ	60.5	**62.1**	62.0
HellaSwag	**45.4**	43.4	42.8
PIQA	69.2	68.7	**69.3**
WinoGrande	55.3	**56.2**	54.6

Table 7: Accuracy of GPT-2 Large model variants on BoolQ, HellaSwag, PIQA, and WinoGrande.

Mechanistic interpretability analysis

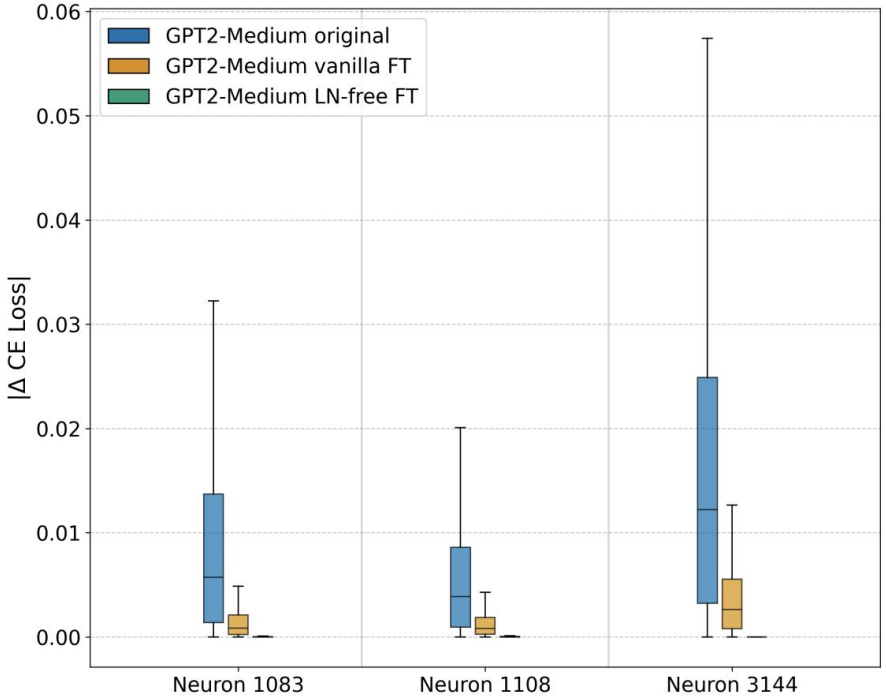
- First-position token no longer special
- Confidence neurons stop working
- Direct Logit Attribution gives direct effect
- Attribution patching error shows no improvement

First-position tokens no longer special



Confidence neurons inactive

- Confidence neuron mean ablation no longer impacts CE loss in LN-free models



Direct logit attribution gives direct effect

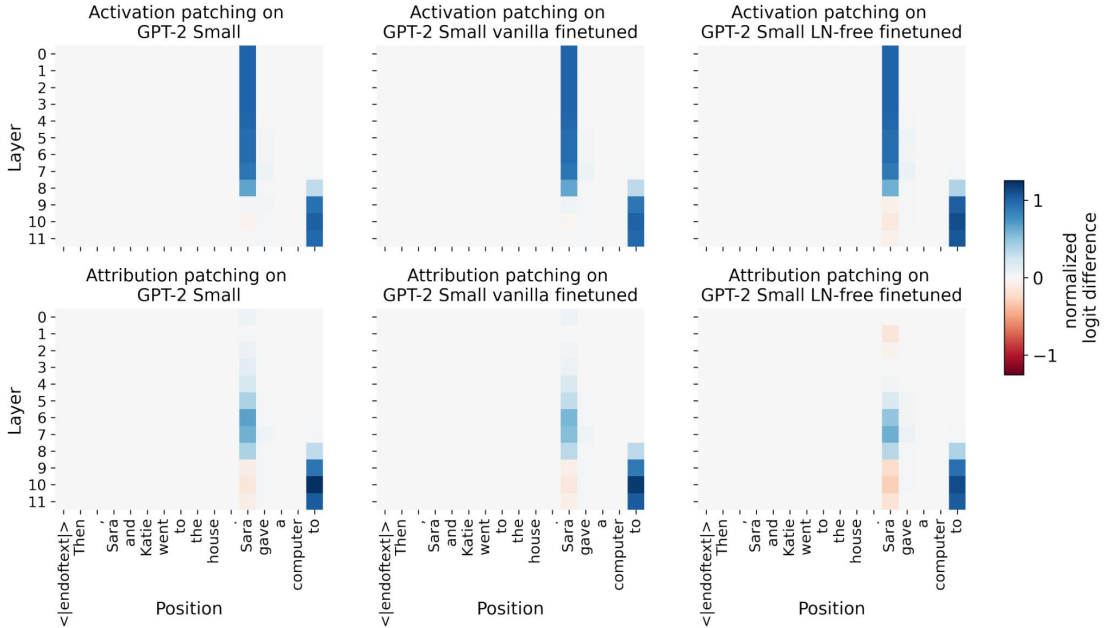
- DLA approximates direct effect of component in residual stream
- LayerNorm removal makes approximation exact as measured by Mean Absolute Error over 1000 examples.

$$\mathbf{DLA}(c) = \mathbf{LN}_{\text{cached}}(c) \cdot W_U,$$

$$\mathbf{DE}(c) = \mathbf{LN}(r) \cdot W_U - \mathbf{LN}(r - c) \cdot W_U,$$

Attribution patching is no more accurate

- Scalable approximation
- LN-free models do not improve attribution patching accuracy



Conclusion

- GPT-2 does not need LayerNorm at Inference Time
- You are welcome to try our models in your experiments
- Follow up work
  - Does removal generalize beyond GPT-2 XL (and Pythia)?
  - Why are LN-free models overconfident?
  - What are the causes of instability during removal?
- Speak to us at the poster session Thu, Apr 23 11:15am - 1:45pm PDT