TRANSFORMERS ALSO NEED STATE: TRAINING-FREE TEMPORAL RESIDUAL CONNECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Residual connections made deep networks trainable, yet standard transformers leave them unused across time. Parallel training optimizes an independent next-token loss at each position, while the model is never asked to generate as a continuous process. The residual stream accumulates layer by layer but is discarded once a token is emitted, surviving only as keys and values for indirect addressing, never as state in the Long Short-Term Memory sense. Multi-step autoregressive generation thus lies outside training. Without a feedback loop, deterministic decoding collapses into repetition; temperature sampling survives ingeniously, but as a compromise. We propose Training-Free Temporal Residual Connections (TF-TRC): the previous step's deep hidden representation is added to the current token embedding at inference, scaled to a small fraction of the embedding norm. With no architectural change and negligible inference overhead, transformers then carry state across time, exploiting the residual stream, although never trained to do so. Across eleven pretrained models from four families and multiple generations, paired bootstrap (over the original eight) shows significant gains on four: filler runs (a degeneration metric) shrink by a third while burst frequency stays unchanged. Same-energy random vectors are harmful, as is injection into Mamba; injection into an AWD-LSTM is inert; transformers and the linear-attention layers of Qwen3.5 benefit. The delimiting counterexample is an xLSTM — it carries state by construction, yet degenerates under sampling, and the same injections turn harmful on it. Only the additive form works; fully replacing the embedding collapses. The gain requires sampling and turns negative under greedy decoding and at low temperature. Its optimum measures how much state-carrying a model is missing.

1 INTRODUCTION

Residual connections made deep networks trainable (He et al., 2016); in a transformer they run layer by layer but never across time. Training never asks the model to generate. Parallel prediction optimizes an independent next-token loss at every position, each conditioned on a human-written prefix, so the model learns to continue other text and never its own; RNN training met this gap as exposure bias (Bengio et al., 2015; Ranzato et al., 2016; Lamb et al., 2016), and the transformer inherits it structurally, since the residual stream is discarded once a token is emitted and survives only as keys and values for indirect addressing (Elhage et al., 2021), never as state in the Long Short-Term Memory sense (Hochreiter & Schmidhuber, 1997). Multi-step autoregressive generation thus lies outside training. Deterministic decoding collapses into repetition (Holtzman et al., 2020); the accounts run through the loss (Fu et al., 2021), the data (Li et al., 2023c), and the attention (Finlayson et al., 2024); temperature sampling survives as a compromise. All of these act at the readout or in training. The state fed back into the model is what has been left alone.

We invert the training recipe. Learning accumulated the representation layer by layer while continuing a true prefix; generation hands that finished representation back. Training-free temporal residual connections (TF-TRC) add the previous step's deep hidden representation to the current input at inference, scaled to a small fraction of its norm; no architectural change, negligible overhead, and the additive form is the only one that survives, since fully replacing the embedding collapses. With this one addition the transformer carries state across time, through the residual stream it already has, although never trained to do so.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: TF-TRC configurations ($\hat{h} = h/\|h\|$, $\hat{d} = d/\|d\|$).

	signal h_{t-1}	signal d_{t-1}
input interface h_0	$h_t + \alpha\|h_t\|\,\hat{h}_{t-1}$	$h_t + \alpha\|h_t\|\,\hat{d}_{t-1}$
deep interface h_l	$h_t + \alpha\, h_{t-1}$	$h_t + \alpha\, d_{t-1}$

Across eleven pretrained models from four families and multiple generations, paired bootstrap (over the original eight) shows significant gains on four: filler runs shrink by a third while burst frequency stays unchanged, and the cross-model default gain 0.22, found on small models, is optimal again at both the input and the deep interface of a 27B model. The controls delimit where the effect lives: same-energy random vectors are harmful, injection into Mamba is harmful, injection into an AWD-LSTM is inert, and xLSTM, recurrent by construction, still degenerates under sampling with the same injections turning harmful on it, so recurrence alone does not save generation and the missing carrier is the residual stream. The mechanism is a readout weighting rather than added memory: the patch replaces the model's default with the trajectory's own history at low-confidence positions, the previous-position head L1_H0 gains $+44\%$ attention, and no information lifetime is extended. The boundary is sharp: the gain requires sampling, greedy locks are absorbing states, and external judges trained on aligned text fail or reverse on our unaligned baselines. Its optimum measures how much state-carrying a model is missing.

2 Temporal Residual Connections

Training-free temporal residual connections (TF-TRC) add the previous step's deep hidden representation to the current hidden representation at inference, scaled to a small fraction of its norm. With no architectural change and negligible inference overhead, the transformer then carries state across time, exploiting the residual stream, although never trained to do so. We present TF-TRC here; T-TRC is left to future work.

We define h_t as the hidden state of the current token at the selected interface. The interface can be the input stream, $h_0 = e_t + \mathrm{pe}_t$ (rotary models: $h_0 = e_t$), or a deep layer, the output of block l (the l-th entry of `output_hidden_states`). We split the models by positional encoding: the learned-PE family (GPT-2) uses `wpe`, with basis $\|e + \mathrm{pe}\|$; the RoPE family uses the embedding alone, with basis $\|e\|$. In both cases h_0 is the full input stream.

$\tilde{h}_t$ is the injection target, and the baseline takes the original h_t. The core update is

$$\tilde{h}_t \;=\; h_t \;+\; \alpha\, s_{t-1}, \tag{1}$$

where s_{t-1} is the signal from the previous token at the source layer l, either the full hidden state h_{t-1} or its windowed residual $d_{t-1} = h_{t-1} - m(W)$ (5-token window, $m(W) = \frac{1}{5}\sum_{j=0}^{4} h_{t-1-j}$). At the input interface the signal enters normalized, $\|h_t\|\,\hat{s}_{t-1}$; at the deep interface the signal itself enters.

Injection starts at the first generated token, with the prompt prefix untouched (the prefix can likewise be processed under TF-TRC step by step; parallel injection is out of scope). Each step costs one $O(d_{\mathrm{model}})$ addition. At the input interface the injection enters the stream directly; at the deep interface it modifies the value in transit from block l to block $l{+}1$, which runs once.

Since the signal and the interface each take two values, we arrange them into the 2×2 grid of Table 1. The two interfaces normalize differently, so we use the dimensionless α as a percentage to align the two injection scales; the two α's coincide in their working range.

Per model, TF-TRC carries three configuration degrees of freedom: the source layer l, the gain α, and the signal (h_{t-1} or d_{t-1}). We set all three once per model by a short dose scan, as one would tune an adapter or a LoRA rank. Across the eleven models the working range of α is 0.05–0.5, and 0.22 works as a cross-model default.

Qwen3.5-architecture models alternate one full-attention block with three linear-attention blocks, so a layer sweep crosses both block types. In the hybrid model we test, layer selection follows the

block type rather than the depth: at matched depth a linear-attention source improves generation, while a full-attention source locks it into repetition.

3 EXPERIMENTS

The eleven models are GPT-2 Small/Medium/Large, Pythia-70m/160m/410m, SmolLM2, Qwen2.5-0.5B and 1.5B, Qwen3.5-0.8B, and a 27B Qwen. We generate 512 tokens per sample at temperature 0.9 with top-k 50, on ten open-ended prompts and two seeds each, and the prompt prefix is never injected. Metrics are pooled per group with early-EOS texts (shorter than 200 tokens) excluded: filler ratio (mean length of runs of tokens the model rates at $p \geq 0.5$), rep3 (token-level three-gram repetition rate), d2 (word-level distinct-2), and maxrun (longest same-token run).

Paired bootstrap over the eight original models shows significant gains on four, GPT-2 Small at 512 tokens (n=20, diff -0.123, CI95 $[-0.249, -0.004]$), Large at α=0.1 (-0.106), Pythia-410m at α=0.5 (-0.238), and Qwen2.5-0.5B at α=0.22 (-0.366), while SmolLM2 at 0.15 falls one line short (-0.179, P=0.950). The non-significant cases each have a cause of their own: 70m lacks power (point estimate -0.094 with CI ± 0.36), 160m sits in a magnitude valley (-0.044), and Medium hits a ceiling (its baseline filler 1.380 is the lowest of the eight).

What gets cut is the pathological part, not filler in general: at the optimal gain the ignition count is unchanged ($184 \rightarrow 178$ per thousand tokens, -3%) while long ignitions (runs of ≥ 3 filler words) disappear ($26.9 \rightarrow 18.6$, -31%), and one-to-two-word runs increase; overdose at α=0.5 does the opposite, with fewer ignitions but longer stretches (max 308), a freeze rather than a recovery.

The strongest single result is on the 27B model: the same 0.22 is optimal at both interfaces, so an injection entering at the input propagates through 48 blocks and lands as if injected at the deep interface; the residual stream carries the d_t information across layers with fidelity, and the input-interface effect is real rather than an artifact. In numbers, $d_t \times 0.22$ at the deep interface beats the baseline on both rep3 (0.071 vs 0.093) and d2 (0.875 vs 0.847), while full h at the same gain degrades (0.111) and α=0.33 collapses, and at the input interface a seven-point grid around 0.22 reproduces the same optimum with the two interfaces agreeing within one sem; the cross-model default found on Small, 0.5B and 1.5B thus carries to 27B and to both interfaces. The input-interface h sweep over the full range shows two collapse points (0.22 and 0.31), recovery to baseline above 0.32, no effect below 0.15, and maxrun 1.1–1.7 throughout (collapse without a hard lock). At the same 0.22, d_t is optimal and h collapses; the sign at matched dose follows the signal.

Same position, same energy, and the only difference is direction: white noise does nothing under teacher forcing (argmax hit 0.372 vs 0.371; the network washes out vectors it cannot read) while TF-TRC flips 12.7% of argmax. In generation the same-energy random direction is not neutral: resampled at every step it triples the filler ratio ($1.40 \rightarrow 4.35$ on Small). A random orthogonal rotation of the signal kills the effect, whitening kills it, and projection onto the top-64 embedding subspace performs like full h, so alignment per dimension and the original spectral profile both matter; fully replacing the embedding collapses; and under greedy decoding every variant locks into repetition (fill 0.93–0.955, max run ≈ 482, locked within 30–60 tokens and never escaped, delayed 10–14 tokens at best), with the gain turning negative below temperature 0.8.

Injection into Mamba is harmful, injection into an AWD-LSTM is inert, and within one hybrid model the two block types respond with opposite signs at matched depth (a full-attention source at layer 19 degrades with filler $+48\%$ and a repetition lock, while linear-attention sources at layers 20/22 improve across the board). The delimiting case is xLSTM-7B: its sampling baseline is itself deeply degenerate (rep3 0.774, vs ≈ 0.09 for the transformers) and greedy collapses completely (0.974, maxrun 200); injecting h at 0.20 improves rep3 and maxrun but lowers d2 (not the TF-TRC signature), 0.30 doubles maxrun, and the windowed residual d_t is uniformly harmful (rep3 $+7\%$, maxrun 2–3$\times$, d2 -45%). Recurrence alone does not save generation, and the same operation that helps transformers turns harmful here.

4 ANALYSIS

The pathology lives in the feedback, not in the operator. Feeding ten human-written passages through the model with zero layers of feedback (teacher forcing on a true prefix) produces no repe-

tition at all (max run 6, vs 482 under greedy self-feedback; d2 0.859), reading out argmax without feeding it back is equally clean, and one layer of self-feedback already fills 22.8%; conversely, reading out without feeding back breaks coherence, because each step conditions on the true prefix rather than on the model's own past, so coherence and in-distribution-ness are mutually exclusive on the discrete realization channel and any readout-without-feedback scheme inherits this.

Degeneration itself is the retreat to the model's default at low-confidence positions, and the patch acts exactly there: under teacher forcing, injection flips 12.7% of positions (16.5% where the top-1 probability is below half, 0.6% above) while the argmax hit rate on human text is unchanged, and the sampled-token statistics stay glued to the human statistics. TF-TRC replaces the model's default with the trajectory's own history. Attention of the previous-position head L1_H0 rises $0.278 \rightarrow 0.401$ ($+44\%$, same direction in 96.3% of steps, 10/10 texts, the largest delta among all 270 layer-head pairs at $7.5\times$ the runner-up), uncorrelated with KL and strongest inside filler stretches, so the mechanism acts where the behavior acts. Probes on trajectories with and without injection show no difference in information decay (≤ 0.015 at all lags, zero beyond $k \geq 8$): TF-TRC does not extend the lifetime of any information but raises the participation volume of the previous step's finished state (a readout weighting, not added memory).

At the deep interface αh decomposes into three operators: the cross-text mean is a temperature operator (OLS logit slope $0.973 \rightarrow 0.668$ with α, linear at 0.28 K/α), the windowed skeleton (the window mean minus its cross-text average) pulls toward contextual collocations (slope 1.005, the only operator with negative correlation to $-\log p$), and the windowed residual d_t is a pure direction operator (zero temperature, zero information preference, cosine 0.80 with the injected vector, the most amplified by the network). Injecting h therefore acts through all three operators at once, while injecting d_t acts through direction alone. In the behavioral dose range (0.05–0.5 at the input interface) the temperature increment is only $+0.001$–0.007 K, so temperature is the overdose pathology channel ($\alpha \geq 0.5$ leaks, α=2 flattens all three) rather than the working channel, and d_t is the cleaner signal. The layer write carries the gain while the input echo inside h is harmful (injecting only the echo component degrades filler, replicated on two batches): the window reference is necessary, the self-echo is poison.

A single injected pulse decays monotonically ($\|\Delta h\|/\|h\|$ of 1.5% at k=1, 0.73% at k=5, 0.07% at k=100, a fast core with half-life ~ 5 steps plus a slow tail), yet closed-loop the same pulse forks the token trajectory within 1–60 steps in 7 of 8 seeds: dissipative hidden dynamics plus sampling sensitivity to small logit shifts, short-range closed form and long-range forked.

Membership, not energy, is the validity condition. White noise is washed out, AWD-LSTM's h is a gate output rather than a stream member (and TF-TRC is inert there), and within Qwen3.5 a full-attention block's h is redundant with the KV channel (injection acts as perturbation plus repeated supply of an existing pattern, a positive feedback into repetition) while a linear-attention block's h is a recurrent state summary outside KV addressing and injection is complementary supply; the criterion generalizes across all three architectures, with the sign decided by the redundancy between the signal's information form and the architecture's existing access channel, and the xLSTM result lands here because its saturated gates put h on the wrong side of the membership line.

Under greedy decoding the lock is an absorbing state: once locked, the injected residual is periodic and in phase with the repetition, so self-sourced feedback cannot escape, and four variants (absolute value, minus realization, minus window mean, minus mean) all fail. What sampling supplies is exploration (basin hopping), not continuity; continuity and exploration are two independent missing components, and TF-TRC extends the working range of sampling but cannot replace it.

5 LIMITS

Sampling is a working condition: under greedy decoding every variant fails, and the gain turns negative below temperature 0.8, so TF-TRC cannot replace sampling. α and temperature are nearly exchangeable on one-dimensional matched statistics but not in mechanism; the matching function exists and is monotone with negative slope (T$0.8 \rightarrow \alpha \approx 0.9$, T$0.9 \rightarrow \alpha \approx 0.47$, T$0.95 \rightarrow \alpha \approx 0.2$, T1.0 leaves the physical region), so the matching surface absorbs one degree of freedom, while temperature rescales the global shape and α selectively flattens low-confidence positions and carries

direction; the surface was calibrated on ten literary and speech passages with one small model, and multi-genre, multi-model recalibration is open.

No static layer-selection criterion survives (the mean-similarity curve criterion fails on Qwen with two counterexamples), so we select the source layer by a short empirical sweep, and no closed-form α exists either: the energy-share construct $\alpha = x/(1+x)$ spans only 0.328–0.402 across twelve models yet overshoots the behavioral optimum by $1.9\times$, so the gain is set by dose scan. External judges we do not trust: judges trained on ChatGPT-grade aligned text fail or reverse on our GPT-2-grade unaligned baselines, the writing-quality reward model has a repetition blind spot (it scores the collapsed repetition groups highest of all), and the repaired detector shows no discrimination across groups (0.89–0.99). The evidence grades differ: the 0.8B block-type contrast and the xLSTM grid are single-dose, $n{=}20$ directional evidence, and the 27B same-optimum result is pattern-level (paired t of $-0.52/-0.39$ at 0.22, not significant), which we claim as a mode rather than a significance.

6 Related Work

Degeneration under autoregressive sampling is well documented (Holtzman et al., 2020), with theoretical accounts (Fu et al., 2021; Li et al., 2023c) and causal follow-ups (Finlayson et al., 2024); truncation and typical sampling reshape the readout (Hewitt et al., 2022; Meister et al., 2023), contrastive decoding contrasts models (Su et al., 2022; Li et al., 2023b), and unlikelihood training penalizes repetition in the objective (Welleck et al., 2020). All of these act at the readout or in training, while TF-TRC acts on the state fed back into the model, requires no training, and composes with any of these decoders. The residual-stream view (Elhage et al., 2021), the logit lens (nostalgebraist, 2020; Belrose et al., 2023) and linear probes (Alain & Bengio, 2017) read the stream layer by layer, whereas we use it as a channel: a bypass that already exists across layers, reused across time at inference.

Prior attempts at temporal reuse modify architecture or training, with Transformer-XL caching hidden states across segments (Dai et al., 2019), compressive transformers compressing them (Rae et al., 2020), block-recurrent and recurrent-memory transformers adding trained recurrence (Hutchins et al., 2022; Bulatov et al., 2022), and universal transformers looping in depth (Dehghani et al., 2019); TF-TRC needs neither. Inference-time interventions steer content, whether truthfulness (Li et al., 2023a), behavior (Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023), or factual readout by layer contrast (Chuang et al., 2024), while TF-TRC feeds back the model's own previous state, and the signal's membership in the stream is exactly what the random-vector control shows to be necessary.

Recurrent and hybrid architectures carry state natively, from LSTM (Hochreiter & Schmidhuber, 1997) and AWD-LSTM (Merity et al., 2018) through S4 and Mamba (Gu et al., 2022; Gu & Dao, 2024), RWKV (Peng et al., 2023), xLSTM (Beck et al., 2024) and linear attention (Katharopoulos et al., 2020) to hybrids (De et al., 2024; Lieber et al., 2024). Our architecture results (Mamba harmful, AWD-LSTM inert, xLSTM the delimiting counterexample, linear-attention blocks within a hybrid the beneficiary) delimit where an additive temporal residual helps: not where state already exists, but where the residual stream is the missing carrier.

AI Use Statement

The author wrote all content in Chinese and made all experimental design decisions and judgment calls; the author also did the literature retrieval, wrote, ran, and analyzed the experiment scripts, and worked on the translation. AI assistance covered English translation and polishing, literature retrieval, and the writing, running, and analysis of experiment scripts.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop*, 2017.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. In *NeurIPS*, 2024.

Nora Belrose, Zachary Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. In *arXiv:2303.08112*, 2023.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *NeurIPS*, 2015.

Aydar Bulatov, Yuri Kuratov, and Mikhail S. Burtsev. Recurrent memory transformer. In *NeurIPS*, 2022.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *ICLR*, 2024.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *ACL*, 2019.

Soham De, Jade Copet, Jirí Mikuła, et al. Recurrentgemma: Moving past transformers for efficient large language models. *arXiv:2404.10190*, 2024.

Mostafa Dehghani, Stephan Sepassi, Lukasz Kaiser, and Geoffrey Hinton. Universal transformers. In *ICLR*, 2019.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Matthew Finlayson, Chandan Reddy, Gal Sharir, Adam Kalai, and Honglak Lee. Closing the curious case of neural text degeneration. In *ICLR*, 2024.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. A theoretical analysis of the repetition problem in text generation. In *AAAI*, 2021.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *COLM*, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *ICLR*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

John Hewitt, Christopher D. Manning, and Percy Liang. Truncation sampling as language model desmoothing. In *Findings of EMNLP*, 2022.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *ICLR*, 2020.

DeLesley Hutchins, Taylor Balch, and Markus Riedel. Block-recurrent transformers. In *NeurIPS*, 2022.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *ICML*, 2020.

Alex Lamb, Anirudh Goyal, Ying Zhang, Saizheng Zhang, Aaron Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. In *NeurIPS*, 2016.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *NeurIPS*, 2023a.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *ACL*, 2023b.

Yixiao Li, Mohit Iyyer, Luke Zettlemoyer, and Jason Eisner. Repetition in repetition out: Towards understanding neural text degeneration from the data perspective. In *AAAI*, 2023c.

Opher Lieber, Barak Lenz, Hofit Bata, et al. Jamba: A hybrid transformer-mamba language model. *arXiv:2403.19887*, 2024.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally typical sampling. *TACL*, 2023.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing lstm language models. In *ICLR*, 2018.

nostalgebraist. Interpreting GPT: the logit lens. LessWrong, 2020.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, et al. Rwkv: Reinventing rnns for the transformer era. In *Findings of EMNLP*, 2023.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *ICLR*, 2020.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *ICLR*, 2016.

Nina Rimsky, Nick Turney, Neel Nanda, et al. Steering llama 2 via contrastive activation addition. In *ACL*, 2024.

Yixiao Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. A contrastive framework for neural text generation. In *NeurIPS*, 2022.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv:2308.10248*, 2023.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *ICLR*, 2020.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv:2310.01405*, 2023.