Equivalence of Context and Parameter Updates in Modern Transformer Blocks

A constructive view of token-specific implicit weight patches

Paper by Adrian Goldwaser, Michael Munn, Javier Gonzalvo, Benoit Dherin

arXiv:2511.17864v3 | Paper discussion, 26 October 2026

Can a prompt be represented as a weight patch?

The paper's question

For a fixed query token, can a model with reduced context and updated parameters produce the same output as the original model with full context?

$$T_{\Theta'}(C \setminus Y, \mathbf{x}) \stackrel{?}{=} T_{\Theta}(C, \mathbf{x})$$

- Y: the context to absorb; Θ': patches to existing parameters.
- Equality is constructed for the current query and generation history.

Paper, Secs. 1 and 3, PDF pp. 1–3.

Modern blocks remove the bias used by the earlier proof

Earlier vanilla block

- Rank-one input matrix patch aligns the MLP activation.
- Output bias absorbs the attention residual shift.

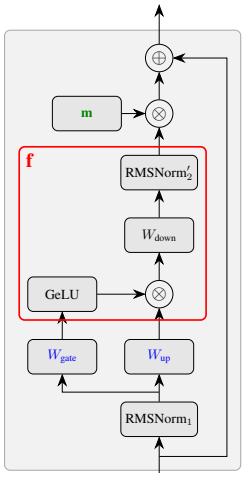
Modern LLM blocks

- Gated MLPs, pre/post normalization, often no MLP bias.
- Need another output parameter to carry the residual shift.

The paper derives the replacement for Gemma, then isolates the general condition.

Paper, Secs. 1–2, PDF pp. 1–2.

Gemma has two places where a patch can act



1. $W_{\text{gate}}, W_{\text{up}}$ control the inputs to the gated MLP.
2. The trainable post-RMSNorm scale **m** controls its final output, coordinate by coordinate.

Division of labor

Input patches reproduce the internal MLP state; output-scale patch restores the missing residual contribution.

Original paper Fig. 1, PDF p. 3; cropped from the paper PDF.

Notation: isolate the change caused by context

$$\mathbf{v}_C = A(C, \mathbf{x}), \qquad \mathbf{v} = A(C \setminus Y, \mathbf{x}), \qquad \Delta\mathbf{v} = \mathbf{v}_C - \mathbf{v}$$

$$\mathbf{z}_C = N_{\mathrm{RMS}}(\mathbf{v}_C), \qquad \mathbf{z} = N_{\mathrm{RMS}}(\mathbf{v})$$

Gemma-style residual MLP output

$$T(C, \mathbf{x}) = \mathbf{v}_C + \mathbf{m} \odot f(W_{\mathrm{gate}}\mathbf{z}_C, W_{\mathrm{up}}\mathbf{z}_C)$$

f includes gating, W_{down}, and post-RMSNorm *without* its scale $\mathbf{m}$.

Paper, Sec. 3.1, Eq. (1), PDF pp. 2–3.

A rank-one patch aligns each input projection

Construction for $i \in \{\mathrm{gate}, \mathrm{up}\}$

$$\Delta W_i = \frac{W_i(\mathbf{z}_C - \mathbf{z})\mathbf{z}^\top}{\|\mathbf{z}\|_2^2}, \qquad \mathbf{z} \neq \mathbf{0}$$

$$(W_i + \Delta W_i)\mathbf{z} = W_i\mathbf{z} + W_i(\mathbf{z}_C - \mathbf{z}) = W_i\mathbf{z}_C.$$

- The gate and up activations now match their full-context targets.
- ΔW_i is an outer product, hence has rank at most one.

Paper, Theorem 1 and Eqs. (2)–(3), PDF p. 3.

The output scale restores the missing residual

Let $\mathbf{h} = f(W_{\text{gate}}\mathbf{z}_C, W_{\text{up}}\mathbf{z}_C)$ after the input patches.

Output correction

$$\Delta\mathbf{m} = \Delta\mathbf{v} \oslash \mathbf{h}, \qquad h_j \neq 0 \text{ for every coordinate } j.$$

$$\underbrace{\mathbf{v} + (\mathbf{m} + \Delta\mathbf{m}) \odot \mathbf{h}}_{\text{reduced context + patches}} = \underbrace{\mathbf{v}_C + \mathbf{m} \odot \mathbf{h}}_{\text{full context}}.$$

The equality is algebraically exact when its nonzero conditions hold.

Paper, Theorem 1 and Eq. (4), PDF p. 3.

The unified theorem needs two controllable functions

$$T(C, \mathbf{x}) = A(C, \mathbf{x}) + g(f(A(C, \mathbf{x}); \theta_f); \theta_g)$$

1. **Input controllability:** update θ_f so the reduced-context input gives the full-context inner activation.
2. **Output controllability:** update θ_g so the outer branch adds any required $\Delta\mathbf{v}$.

Sufficient condition

When both properties hold, with the required activations nonzero, a context-equivalent patch exists for the residual block.

Paper, Definitions 1–2 and Theorem 3, PDF pp. 7–8.

Different architectures offer different output controls

Outer control	Patch that adds $\Delta\mathbf{v}$	Condition
Bias $\mathbf{b}$	$\Delta\mathbf{b} = \Delta\mathbf{v}$	trainable bias
Output matrix W'	$\Delta W' = \Delta\mathbf{v}\,\mathbf{h}^\top / \|\mathbf{h}\|^2$	$\mathbf{h} \neq 0$
RMSNorm scale $\mathbf{m}$	$\Delta\mathbf{m} = \Delta\mathbf{v} \oslash \mathbf{h}$	every $h_j \neq 0$
MoE experts	controllable experts share correction	$S \neq 0$

Gated and pre-normalized inputs remain controllable through their input matrices.

Paper, Table 1 and Appendix A, PDF pp. 8, 12–13.

A fixed post-RMSNorm scale gives a counterexample

Geometric bound

With fixed $\mathbf{m}$, the post-RMSNorm output lies in a bounded set:

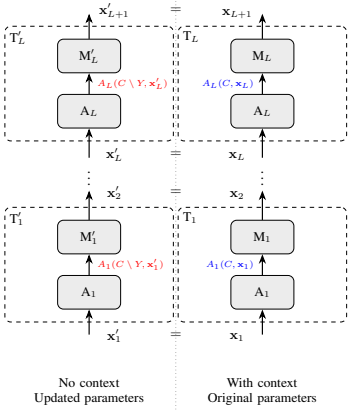
$$\|g(\mathbf{z}') - g(\mathbf{z})\|_2 \leq 2\sqrt{d}\,\|\mathbf{m}\|_\infty.$$

- The attention-induced $\Delta\mathbf{v}$ can exceed this bound.
- Then no change upstream of the fixed normalization can compensate it.

Output controllability has a real architectural boundary.

Paper, Sec. 5.1 and Appendix D, PDF pp. 9, 19–20.

The single-block result composes layer by layer



1. Start from the same query-token representation.
2. If layer l inputs match, patch layer l to match its outputs.
3. The common output becomes the input to layer $l + 1$.

Induction gives equality at the final layer.

Paper, Fig. 2 and Theorem 2, PDF p. 4; cropped from the paper PDF.

The algorithm records targets, then patches sequentially

1. **Full-context pass:** save the target activation after each layer.
2. **Reduced-context pass:** at layer l, compute its patch against the saved target.
3. Forward the layer's *actual patched output* to layer $l + 1$.

During autoregressive generation

Recompute patches for each next-token query because the generation history changes.

Paper, Sec. 3.3, Algorithm 1 and Sec. 4.1, PDF pp. 4–6.

Small output coordinates destabilize the direct Gemma patch

$$\Delta m_j = \frac{\Delta v_j}{h_j}$$

- A small h_j can magnify the scale patch and rounding error.
- The ideal equality from Theorem 1 can drift in `bfloat16`.
- Figure 3 exhibits token mismatches for the direct low-precision update.

Design response

Move most of the correction into W_{down}; use $\mathbf{m}$ for the remaining error.

Paper, Sec. 4.2 and Appendix B, PDF pp. 6, 13–15.

Stable Gemma update: matrix first, scale second

1. Set target output $\mathbf{g} = \Delta\mathbf{v} + \mathbf{m} \odot N(\mathbf{h}_{\text{down},C})$.
2. Find $\mathbf{h}_{\text{target}}$ with a fixed RMS value that best fits $\mathbf{g}$ after scaled RMSNorm.
3. Patch W_{down} to reach $\mathbf{h}_{\text{target}}$; patch $\mathbf{m}$ for the residual.

$$\Delta W_{\text{down}} = \frac{(\mathbf{h}_{\text{target}} - \mathbf{h}_{\text{down},C})\mathbf{h}_{\text{gated},C}^{\top}}{\|\mathbf{h}_{\text{gated},C}\|^2}, \qquad \Delta\mathbf{m} = \mathbf{r} \oslash N(\mathbf{h}_{\text{target}}).$$

Paper, Appendix B.1–B.2, PDF pp. 13–15.

Evaluation compares identical next-token histories

Two generation paths

Original model with full context vs. patched model with no explicit context.

- Recompute the patch at each token, including the prompt and prior generated tokens.
- If argmax tokens differ, continue both comparisons on the baseline history.
- Metrics: token match ↑, L_∞ logit difference ↓, TVD ↓.

$$\mathrm{TVD}(p, q) = \tfrac{1}{2}\|p - q\|_1.$$

Paper, Secs. 4.1–4.2, PDF pp. 5–6.

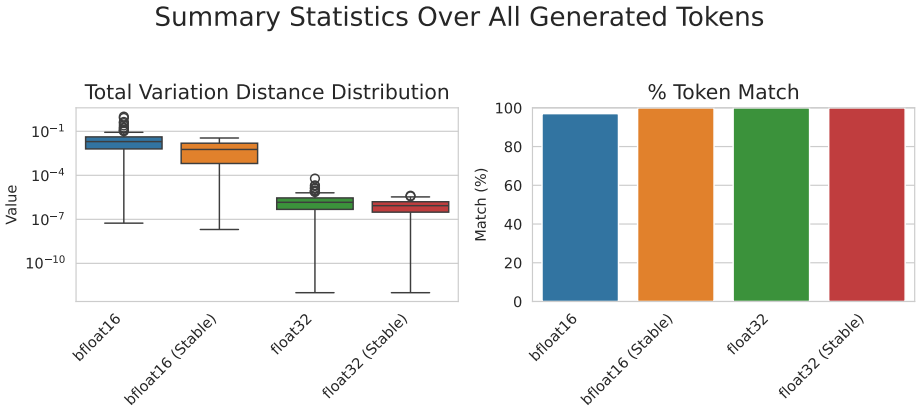
Gemma text matching depends strongly on precision

Gemma 3 1B setting	Token match	Distribution comparison
`float32`, direct	100%	nearly identical
`bfloat16`, direct	87.5%	clear logit / TVD drift
`bfloat16`, stable	100%	drift reduced

The 87.5% figure is the main text's low-precision case; Figure 4 pools five text generations.

Paper, Sec. 4.2, Figs. 3–4, PDF pp. 5–7.

Across five prompts, stable updates recover token match

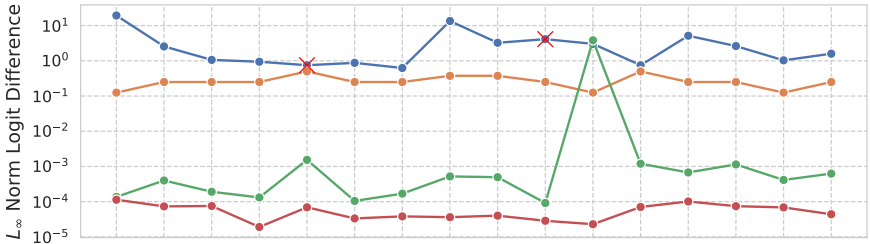


Original paper Fig. 4, PDF p. 7. Left: TVD (log scale; lower is better). Right: token match (higher is better).

The image-context example follows the same pattern



Prompt: <start_of_turn>user <start_of_image>
What is this image? Answer in 10 words:
<end_of_turn><start_of_turn>model



Gemma 3 4B: direct `bfloat16` (blue) has token mismatches; the stable update (orange) reduces logit drift. Both `float32` variants (green/red) stay close.

Original paper Fig. 5, top panel cropped, and Sec. 4.2, PDF p. 7.

Falcon needs an outer matrix patch, not a scale patch

Architecture

Falcon's tested block lacks Gemma's post-normalization scale, so its outer output matrix absorbs $\Delta\mathbf{v}$.

Observed fidelity

Across five prompts, both `bfloat16` and `float32` achieve 100% token match; `float32` has smaller TVD.

$$\Delta W_{\text{out}} = \frac{\Delta\mathbf{v}\,\mathbf{h}^{\top}}{\|\mathbf{h}\|^{2}}$$

Paper, Table 1 and Appendix C.3, Fig. 9, PDF pp. 8, 17–18.

Ablation: distributing patches across layers helps

- Updating only the last layer causes numerical problems in the reported sweep.
- With a few adjustable layers, token accuracy is usually restored.
- Starting earlier lowers the stable scheme's scale-vector update norm.

Interpretation

Existence of a patch leaves many valid constructions; placement and parameter choice determine finite-precision behavior.

Paper, Appendix C.4–C.5 and Figs. 11–13, PDF pp. 17–18, 23–25.

The result is local to a query and its history

1. **Token-specific:** changing the next-token history changes the target activations and required patches.
2. **Conditional:** input and output controls need the stated nonzero activations.
3. **Numerical:** small divisors make direct low-precision evaluation unstable.

Reading of the claim

The construction expresses the context effect as parameter changes at one query; the experiments measure how accurately those changes reproduce logits.

Paper, Sec. 6, Appendix B, PDF pp. 9, 13–15.

Take-home messages

1. Input projections can absorb a context-induced activation shift using rank-one updates.
2. An output-controllable parameter must also absorb the attention residual shift.
3. Layerwise construction extends the equality to deep models; stable parameter choices matter in low precision.

Controllability is the useful architectural test.

Paper, Theorems 1–3, Secs. 3–6, PDF pp. 3–9.

Paper and discussion pointers

- Goldwaser, Munn, Gonzalvo, Dherin. *Equivalence of Context and Parameter Updates in Modern Transformer Blocks.*
- arXiv:2511.17864v3, revised 6 July 2026.
- Main proof: Theorems 1–3; numerics: Appendix B; experiments: Sec. 4 and Appendix C.

Questions and discussion

Backup: RMSNorm inversion is constrained fitting

For target $\mathbf{g}$ and fixed RMS C, find $\mathbf{x}$ minimizing

$$\left\| \mathbf{m} \odot \frac{\mathbf{x}}{\mathrm{RMS}(\mathbf{x})} - \mathbf{g} \right\|_2^2 \quad \text{subject to} \quad \mathrm{RMS}(\mathbf{x}) = C.$$

With $\mathbf{y} = \mathbf{x}/C$, the stationarity condition gives

$$y_k = \frac{g_k m_k}{m_k^2 - \mu}, \qquad \frac{1}{n} \sum_k \frac{(g_k m_k)^2}{(m_k^2 - \mu)^2} = 1.$$

The paper solves for the scalar μ and uses the result as $\mathbf{h}_{\text{target}}$.

Paper, Appendix B.2, PDF pp. 14–15.

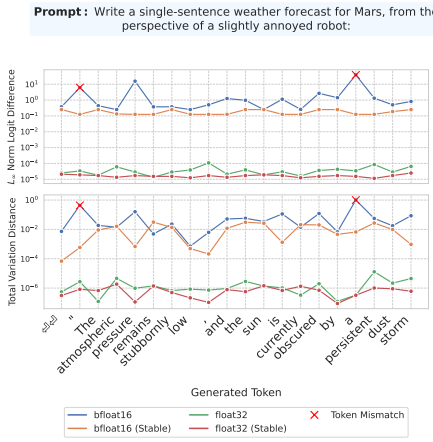
Backup: the gradient-step form is constructed from the patch

$$\mathcal{L}(W) = -\operatorname{Tr}(\Delta^\top W) \quad \Longrightarrow \quad -\nabla_W \mathcal{L} = \Delta.$$

- Choose Δ from the context-equivalent patch and a norm-based step size.
- Summing context increments telescopes to the full patch.
- This supplies an algebraic gradient representation for the desired update.

Paper, Appendix E, PDF pp. 20–22.

Backup: per-token Mars forecast example



Original paper Fig. 3, PDF p. 5. Red crosses mark token mismatches.