What the education-not-advice edit guarantees: eleven theorems checked in Lean

Parker Smith. Medical student, LSU Health Shreveport.
3 November 2026

Purpose. Notebook sections 42 and 43 of the interp-lab project describe an edit to a chat model (Qwen3.5-4B) that makes it explain medicine rather than advise, and they make three "by construction" claims about the linear algebra of that edit. This document states those claims as theorems in an abstract real inner product space and proves them. Every statement below is also a theorem in the Lean 4 proof assistant, checked against the Mathlib library (Lean 4.35.0, Mathlib v4.35.0), in the file `lean/EduProofs/Invariants.lean` of the project repository; the Lean name of each statement is given in parentheses. The proofs are short, and that is the point. The experiment found that the permanent version of the edit preserved exactly what these theorems say it preserves and still broke the model's call-911 answer. The theorems make precise what was guaranteed, so that the result is read correctly: the guarantee was about coordinates of the model's working memory, not about the answer.

Setting. Throughout, V is a real inner product space with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$. In the experiment $V = \mathbb{R}^{2560}$, the width of the model's residual stream, with the ordinary dot product. The theorems hold for any V. Scalars act on vectors by $c \cdot v$.

The removal of a direction

Definition 1 (`removeAlong`)**.** For $u, v \in V$, the removal of u from v is

$$\mathrm{remove}_u(v) = v - \langle v, u \rangle\, u.$$

When $\|u\| = 1$ this is the orthogonal projection of v onto the complement of u. Applied to every matrix W through which the model writes into its residual stream, $W \mapsto (I - uu^{\top})W$, it is the permanent removal of section 42.

Theorem 1 (Claim 1, `inner_removeAlong_self`)**.** *If* $\|u\| = 1$*, then* $\langle \mathrm{remove}_u(v), u \rangle = 0$ *for every* $v \in V$.

Proof. By bilinearity, $\langle v - \langle v, u \rangle u, u \rangle = \langle v, u \rangle - \langle v, u \rangle \langle u, u \rangle$, and $\langle u, u \rangle = \|u\|^2 = 1$. □

Theorem 2 (`removeAlong_of_inner_eq_zero`)**.** *If* $\langle v, u \rangle = 0$*, then* $\mathrm{remove}_u(v) = v$.

Proof. The subtracted term is $\langle v, u \rangle\, u = 0 \cdot u = 0$. □

Theorem 3 (Claim 2, one coordinate; `inner_removeAlong_of_orth`)**.** *If* $\langle u, e \rangle = 0$*, then* $\langle \mathrm{remove}_u(v), e \rangle = \langle v, e \rangle$ *for every* $v \in V$.

Proof. $\langle v - \langle v, u \rangle u, e \rangle = \langle v, e \rangle - \langle v, u \rangle \langle u, e \rangle = \langle v, e \rangle$. □

The guarded direction

Definition 2 (`guard`)**.** Let $E = (e_i)_{i \in I}$ be a finite family in V and $u \in V$. The guarded direction is

$$\mathrm{guard}_E(u) = u - \sum_{i \in I} \langle u, e_i \rangle\, e_i.$$

In the experiment E is an orthonormal family of four vectors spanning the directions along which emergencies differ from routine and same-day cases at layer 16, and u is the advice direction.

Theorem 4 (Claim 2, the construction; `inner_guard_eq_zero`)**.** *If E is orthonormal, that is $\langle e_i, e_j\rangle = 1$ when $i = j$ and 0 otherwise, then $\langle \mathrm{guard}_E(u), e_j\rangle = 0$ for every $j \in I$.*

Proof. $\langle u - \sum_i \langle u, e_i\rangle e_i, e_j\rangle = \langle u, e_j\rangle - \sum_i \langle u, e_i\rangle\langle e_i, e_j\rangle$, and orthonormality collapses the sum to its $i = j$ term, $\langle u, e_j\rangle$. □

Theorem 5 (`inner_smul_guard_eq_zero`)**.** *For every scalar c, $\langle c \cdot \mathrm{guard}_E(u), e_j\rangle = 0$. In particular the normalised guarded direction $u_g = \mathrm{guard}_E(u)/\|\mathrm{guard}_E(u)\|$ is orthogonal to the family.*

Proof. $\langle c \cdot w, e_j\rangle = c\langle w, e_j\rangle = c \cdot 0$. □

Theorem 6 (Claim 2, put together; `inner_removeAlong_guard`)**.** *If E is orthonormal, then for every scalar c, every $v \in V$ and every $j \in I$,*

$$\langle \mathrm{remove}_{c\cdot \mathrm{guard}_E(u)}(v), e_j\rangle = \langle v, e_j\rangle.$$

Proof. Combine the previous theorem with Claim 2 for one coordinate. □

The residual stream

Definition 3 (`stream`)**.** For an embedding $e \in V$, a sequence of block writes $w_0, w_1, \ldots \in V$ and a depth n, the residual stream after n blocks is

$$\mathrm{stream}(e, w, n) = e + \sum_{i<n} w_i.$$

The writes are arbitrary vectors. In a real network each block computes its write from the stream it reads, so the writes change when the stream changes; the theorems below hold for whatever the blocks write.

Theorem 7 (Claim 3, first half; `inner_stream_edited_self`)**.** *If $\|u\| = 1$, then for every embedding e, every sequence of writes w and every depth n,*

$$\langle \mathrm{stream}\big(\mathrm{remove}_u(e),\ (\mathrm{remove}_u(w_i))_i,\ n\big), u\rangle = 0.$$

Proof. The inner product with u is additive, and each term is zero by Claim 1. □

Theorem 8 (Claim 3, second half; `inner_stream_edited_coord`)**.** *If $\langle g, x\rangle = 0$, then for every embedding e, every sequence of writes w and every depth n,*

$$\langle \mathrm{stream}\big(\mathrm{remove}_g(e),\ (\mathrm{remove}_g(w_i))_i,\ n\big), x\rangle = \langle \mathrm{stream}(e, w, n), x\rangle.$$

Proof. The inner product with x is additive, and each term is unchanged by Claim 2 for one coordinate. □

This is the whole of the guarantee behind the permanent removal: with $g = u_g$ and x any member of the emergency family, every emergency coordinate of whatever the blocks write passes through the edit unchanged, at every depth. It says nothing about what the blocks write, and the blocks of the edited model read an edited stream. That is where the call-911 answer was lost.

The switch

Definition 4 (`gated` and `switch`)**.** For a push $s \in V$ and a state $v \in V$, the conditional edit is

$$\mathrm{gated}(b, s, v) = \begin{cases} v - s & \text{if } b \text{ is true,} \\ v & \text{otherwise,} \end{cases}$$

and the switch, for an advice score a, an emergency score m and thresholds τ_a, τ_m, is

$$\mathrm{switch}(a, m, \tau_a, \tau_m) = \big[\tau_a < a \text{ and } m \le \tau_m\big].$$

Theorem 9 (Claim 4; `gated_false`)**.** $\mathrm{gated}(\mathit{false}, s, v) = v$ *for every* $s, v \in V$.

Proof. By the definition. □

Theorem 10 (`switch_eq_false_of_emergency`)**.** *If* $\tau_m < m$*, then* $\mathrm{switch}(a, m, \tau_a, \tau_m)$ *is false, whatever the advice score* a*.*

Proof. The second conjunct $m \le \tau_m$ fails. □

Theorem 11 (`gated_switch_identity_of_emergency`)**.** *If* $\tau_m < m$*, then for every push* s *and state* v*,*

$$\mathrm{gated}\big(\mathrm{switch}(a, m, \tau_a, \tau_m), s, v\big) = v.$$

Proof. Combine the two previous theorems. □

The hypotheses, measured on the saved edit

The theorems take hypotheses: the removed direction has unit norm, the emergency family is orthonormal, the guarded direction is orthogonal to it. The saved vectors are 32-bit floating point numbers, so the hypotheses hold to rounding, not exactly. Table 1 gives the measurements, from `scripts/edu_lean_check.py`. The last two rows measure Claim 3 on the real model after the weight edit in 16-bit floating point (bfloat16): the largest component along u_g at any of the 32 layers is 0.010, on residual vectors of norm 2 to 54, which is the size of bfloat16 rounding.

What is not proved

Nothing above concerns the model's behaviour. Where the switch opens on messages it has not seen, whether the explanation the edited model writes is true, and what two readers will call it are empirical questions, answered in notebook sections 42 and 43 and on the website page. The theorems also do not bound the effect of a removal on the model's output through the nonlinear parts of each block (the normalisation and the attention and feed-forward computations); Claim 3 is a statement about coordinates of the stream, and the experiment shows that the coordinates it protects are not the ones the call-911 decision reads.

Quantity	Value	Theorem
Dimension of the residual stream	2560	
Norm of the guarded direction u_g (hypothesis: 1)	0.9999990	Claims 1 and 3
Largest entry of $E^\top E - I$ (hypothesis: E orthonormal)	1.9×10^{-9}	Claim 2
Largest $\lvert E^\top u_g \rvert$ (hypothesis: 0)	6.9×10^{-8}	Claim 2
$u \cdot e_1$ before the guard	+0.0225	why the guard was needed
Part of u inside the emergency family, as a share of its norm	0.228	
Saved patch vector against the fitted vector, largest difference	0	the saved file is the proved object
Component along u_g after removal, on a random vector of norm about 5045	8×10^{-5}	Claim 1 in 64-bit arithmetic
Change in the E coordinates after that removal	2.7×10^{-6}	Claim 2 in 64-bit arithmetic
Component along u_g in the edited model, largest over 32 layers	0.010	Claim 3 in bfloat16
Same, relative to the residual norm	2.0×10^{-4}	Claim 3 in bfloat16

Table 1. Measurements of the theorems' hypotheses on the saved edit, in 64-bit arithmetic from the saved 32-bit vectors, and of the Claim 3 invariant on the model after the bfloat16 weight edit, at the reply opener of one routine message.