SuperBake: Installing Verified Facts into Transformer Weights by Direct Construction

Albert Ruehlman
AMI Labs
howchunconsulting@gmail.com

July 22, 2026

Abstract

We present SuperBake, a system that installs new factual knowledge into the weights of large language models *without any gradient steps*. Instead of fine-tuning, low-rank adaptation, or retrieval, SuperBake measures how the stock model represents each question and then hand-constructs a small circuit of RUNEs (*RU*ehlman *NE*urons) — exact weights, written directly — in a region appended to the network's MLPs. Each installed fact receives a physical address (layer and slot coordinates), is behaviorally verified across phrasing variants, and is *neutralized* (zeroed) if verification fails, so delivered weights never carry silently broken knowledge. The output is a standard checkpoint that stock inference code loads unchanged, together with a receipt listing every fact, its verification status, and its coordinates.

On a 7,450-row battery over 1,000 facts (Llama-3.1-8B-Instruct), the constructed engine verifies 91.3% of rows overall against 76.9% for a strong masked-SGD baseline — while leaving prose perplexity at the stock baseline, where the SGD baseline more than doubles it. Constructed facts answer *reverse* questions at ∼94% where the trained baseline scores 0.7%: the reversal curse does not apply when the reverse mapping is simply written. Delivered, stock-loadable checkpoints reach 93–97% primary-question recall on Llama-3.1-8B and Qwen2.5-7B. Construction is fast: measured end-to-end bakes run in minutes on one 80 GB GPU, and a reduced pipeline wrote 192 facts into a raw Pythia-6.9B at 14.8 facts per second.

The method was derived from a mechanistic post-mortem of what stochastic gradient descent actually builds when it crams facts into appended neurons. We report those findings: dense fact storage lands in a *coherent sub-threshold population code* living in activation-function leakage — the same signal that damages prose — and several transport and capacity laws that any weight-space editor must respect. SuperBake is the constructive rebuild of that mechanism: the same associative memory, but gated, addressable, verifiable, and harmless to the host model.

1 Introduction

Getting new knowledge *into* a language model's weights is still surprisingly hard. Fine-tuning on new facts damages what the model already knows and, at practical fact densities, its general prose ability; parameter-efficient variants inherit the same interference because they optimize the same objective. Retrieval-augmented generation sidesteps the problem by not writing the weights at all — the knowledge lives in an external index, with its latency, context cost, and failure modes. Knowledge-editing methods such as ROME and MEMIT [2, 3] write weights directly, but they *modify existing* projections via least-squares updates, are evaluated largely on single-hop cloze recall, and recent lifelong-editing evaluations at realistic scale find they fall behind simple retrieval or continued fine-tuning [8].

This paper describes a different route. SuperBake treats a fact as a small *circuit to be engineered*, not a gradient target or a database row. The system (1) appends fresh, initially-silent neuron slots to the model's MLPs, leaving every original weight untouched; (2) measures how the stock model represents each question across phrasings and contexts; (3) writes exact weights for a per-fact circuit — recognition keys, an injected code, answer readouts, spelling chains, and delivery insurance — derived in closed form from those measurements; (4) calibrates the result in a measure-and-correct loop that uses no loss function and no gradients; and (5) *verifies* every fact behaviorally, zeroing the neurons of any fact that fails, before saving a standard checkpoint.

Because every weight is written deliberately, each fact has a legible anatomy: a receipt lists the exact layer-and-slot coordinates of its neurons. Ablate them and the fact is gone; leave them and it survives subsequent fine-tuning better than the model's own pretrained knowledge (§4). Because nothing is optimized against the model's existing behavior, the host is unharmed: across our main runs the prose negative log-likelihood of baked models is equal to — in several runs slightly *better* than — the stock baseline.

Contributions.

1. **A zero-gradient installation method** for factual knowledge in transformer LLMs: per-fact constructed circuits in appended MLP slots, closed-form keys and values, and a gradient-free closed-loop calibrator, delivered as bone-stock checkpoints (§3).

2. **A verification-first pipeline**: phrasing batteries, held-knowledge and fluency referees, and a pruner that guarantees *verified or neutralized* — failed facts degrade to honest ignorance, never to garbage (§3.9).

3. **Results** (§4): 91.3% vs. 76.9% overall against a strong masked-SGD baseline at zero prose cost; delivered checkpoints at 93–97%; dissolution of the reversal curse by construction (∼94% vs. 0.7%); durability under further fine-tuning; minutes-scale bakes.

4. **Mechanistic findings** (§5): measured evidence that SGD stores dense fact overflow in a coherent *sub-threshold* population code carried by activation leakage; transport and capacity laws for signals injected into a transformer's residual stream; and a compendium of design laws (§7) for weight-space engineering.

5. **Composition by construction** (§3.8): one additional constructed neuron per fact gives installed facts *silent multi-hop* behavior (using a fact without being asked for it), exceeding what emerges natively at comparable scale.

We close with an honest boundary (§8): free-flowing *conversational* recall on the bare weights — pronouns, ellipsis, novel context mid-dialogue — is measurably harder than any enumerated battery and remains an open frontier; delivered models disclose exactly which phrasings verified conversationally.

2 Background and Notation

Four standard mechanistic-interpretability facts carry the whole method; readers fluent in the circuits literature can skim.

The residual stream is a shared bus. A decoder block reads its input from, and adds its output to, a per-token vector stream $x_\ell \in \mathbb{R}^d$ [9]. Anything a layer writes is visible to every later layer. A constructed neuron at layer 6 can therefore deposit a message that a constructed reader at layer 25 consumes.

MLP neurons are keyed value-writers. In a SwiGLU MLP the i-th neuron computes

$$\text{out}_i \;=\; \underbrace{\text{silu}(g_i \cdot x)}_{\text{gate}}\;\underbrace{(u_i \cdot x)}_{\text{up}}\;\underbrace{v_i}_{\text{value}}, \tag{1}$$

a two-factor match on the current state followed by the addition of a fixed direction v_i to the stream [1]. Writing a neuron by hand means choosing a key (what it fires on), a second factor (an AND condition), and a value (what it writes). Pythia-style MLPs lack the up path and carry biases; the construction adapts (§6).

Attention heads are transport. A head moves information between positions: QK decides *from where*, OV decides *what* is carried [9]. Under rotary position embeddings [16], the QK match acquires a distance-dependent kernel set by which rotary frequencies the key/query directions occupy — a fact we exploit (§3.6) and are limited by (§7).

Directions read out as token preferences. Projecting a residual state onto unembedding rows scores next-token candidates (the "logit lens" [11]); the logit *gap* between the correct answer token and its strongest competitor at the answer position is the scalar our calibrator drives positive.

That is the entire theoretical load. Everything else in the paper is measurement and construction.

3 The SuperBake Engine

Throughout, we call a hand-written, exact-weight unit of Eq. 1 a *RUNE*: a neuron whose key, AND-condition, and value are chosen rather than trained. A fact is a small circuit of RUNEs, spread across the network's depth.

3.1 Appended capacity and stock delivery

Before any writing, every MLP in the model is expanded by the same number E of zero columns and rows — new neurons whose weights are exactly zero, hence perfectly silent. Uniform expansion matters: the delivered config declares one `intermediate_size`, so the checkpoint loads in stock `transformers`/vLLM with no custom code. Facts are then written *into the delivered geometry*: the states we measure during construction are the states the customer's copy will compute. (An early pipeline that compacted and re-ordered columns at save time taught us this the hard way: reordering thousands of columns changes bf16 accumulation order and flips borderline facts; building in final geometry removes the gap. Save/reload round-trips are exact — verified to four decimals of NLL and row-for-row recall.) Unused slots cost disk, so dead neurons are pruned to zero and the expansion is sized from the battery (∼0.5 GB per 1,000 slots per layer at 8B scale).

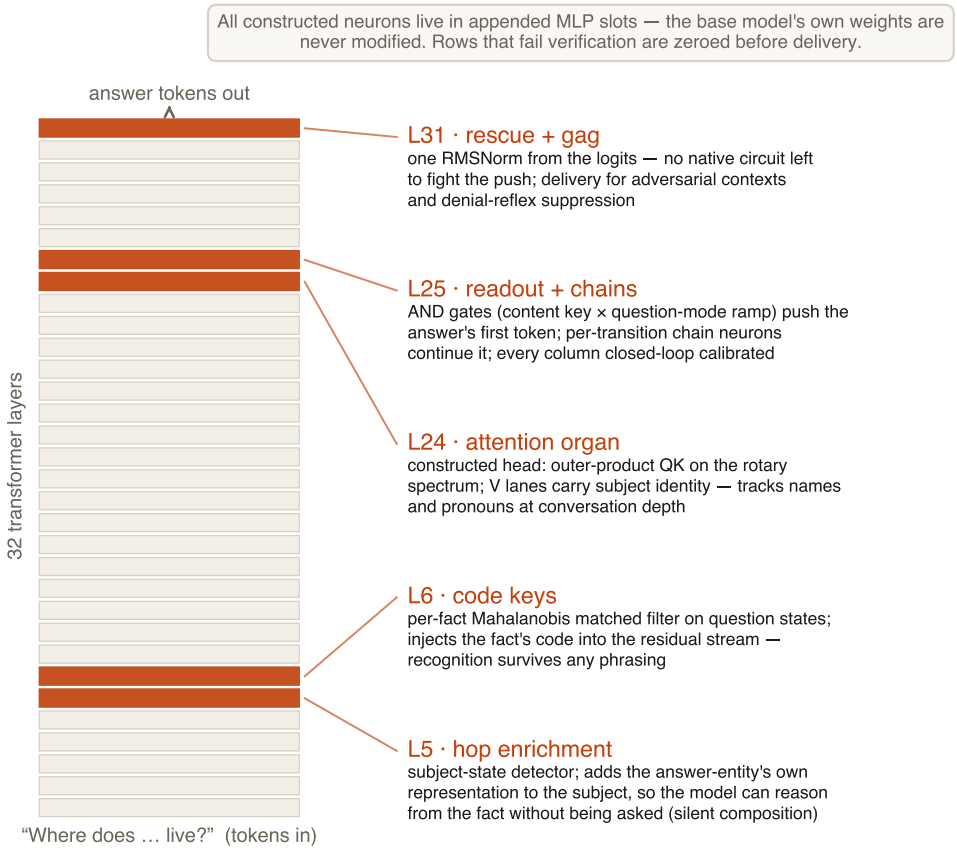

Figure 1: Anatomy of one installed fact on Llama-3.1-8B (32 layers). Each site is a small bank of RUNEs in appended MLP slots; the receipt records every coordinate. Question-recognition keys inject a per-fact code early; readout and chain neurons speak the answer late; an attention organ transports subject identity across conversation positions; a last-layer bank delivers in adversarial contexts.

3.2 Harvest: photographing the stock model

All keys and thresholds are derived from measured states. The engine runs the unmodified model over every question rendered several ways — phrasing variants, casing variants, chat-template and plain renders, and several *lightings*: different batch compositions and conversational prefixes that perturb the states the way deployment will. It also harvests *innocents*: states from generic prose, ordinary chat, and held-out known-fact questions, which become the negative population every gate must clear. Harvesting under multiple lightings and thresholding against the *worst case* across them is what makes keys transfer out of the build context (§6).

3.3 Code keys at layer 6

Each fact f gets one recognition neuron low in the stack. Its key is a Mahalanobis matched filter on the question-end state,

$$k_f \;=\; \Sigma^{-1}\,(\bar{x}_f - \mu), \tag{2}$$

where $\bar{x}_f$ is the mean state over fact f's renders and (μ, Σ) are estimated over the *union* population — all facts' states *and* the innocents. Whitening is mandatory in crowded key space: raw mean keys mostly measure the shared question subspace (own-fact response 0.82 vs. strongest other 4.87

before whitening), and excluding innocents from Σ lets the inverse amplify them. The threshold is set above every other fact's response and every innocent's, with a hard clearance floor. The neuron's value writes a random unit vector c_f — the fact's *code* — into the stream, scaled to a target arrival magnitude.

Two properties do the work. First, the matched filter generalizes over phrasing: any rendering of the question lands near $\bar{x}_f$ after whitening. Second, the injected code manufactures separability downstream: at layer 25, raw question states of different facts overlap badly (median nearest-neighbor margin $\sim$1.1), but with codes present every fact is linearly separable (4,750/4,750 at 1,000-fact scale, median margin 5.8). The code is the address the rest of the circuit keys on.

Born-hard gates. Hand-set thresholds sit in silu's leaky region, and thousands of sub-threshold leaks sum into prose damage (§5). Every constructed gate is therefore *born hard*: scale the gate row and bias by s and divide the value column by s; $\mathrm{silu}(s\,z)/s$ approximates a hard threshold with leak $\leq 0.28/s$, exponentially dead where prose states sit. At s=8–16 constructed banks contribute exactly zero measured prose drift.

3.4 Readout and chains at layer 25

The speaking end is a per-fact *readout* neuron and a bank of *chain* neurons. The readout's gate keys on the code-enriched question state; its up path is a question-mode ramp — a direction that separates "a question is being answered here" from prose, making the neuron a multiplicative AND (Eq. 1) that cannot fire during ordinary text. Its value pushes the answer's first token through the unembedding and *suppresses the measured competitor*: the highest-logit wrong token observed for that fact, with the suppression direction orthogonalized against the answer direction (correlated unembeddings otherwise make suppression drag the answer down with it).

Answers longer than one token are carried by *per-transition* chain neurons — one per (fact, answer-position) — keyed on the mid-answer state and pushing the next token. Three structural rules were each forced by a measured failure: (i) keys are fully per-row, never group-mean (a group-mean key thresholds at its worst phrasing and overshoots its siblings); (ii) rows of the same fact are *allies*, excluded from one another's negative pools (a sibling cross-fire pushes the same token — harmless); (iii) siblings that share a transition each carry $1/n$ of the push, because generation sums them — eleven same-fact rows each sized to close the full gap once blew the residual out of distribution and produced junk argmaxes.

3.5 The closed loop: calibration without gradients

Construction gets facts *near* correct; a measurement loop finishes them. For each fact the engine renders the question, runs the model, and reads the logit gap at the answer position. Failing facts get their own value column rescaled — nobody else's weights move. The step size uses a secant estimate of *transfer*: a push written at layer 25 traverses six blocks and a final RMSNorm, arriving attenuated by a per-fact factor $\hat{f} \approx 0.2$–0.5; measuring (closed gap)/(intended push) and dividing the next intent by $\hat{f}$ converges in a few rounds. Guard rails, each added after a measured pathology: a *backfire guard* (a fact that regresses under a boost gets its column shrunk instead — pure $\hat{f}$ adaptation spirals); a two-backfire freeze; *multi-token suppression* (the last stubborn facts were losing to prefix-fragment tokenizations of their own answers, e.g. `' Le'` for *Le Havre* — the top four wrong tokens are suppressed together); and *front-to-back* chain repair (transition k lives inside the context of k+1, so only the earliest failing transition advances per round; unordered repair oscillates). Chain repairs are confirmed against the verifier's exact render, and readout repair runs

before chain repair — a readout push at the question position propagates through attention into mid-answer states and un-wins transitions repaired first.

3.6 Transport: the attention organ

Everything above keys on states at the question's end — which works until the question lives deep in a conversation, where transcript context shifts those states off the keys. The stable signal, measured across depths, is the state at the *subject's own tokens* (± 0.3 units across six turns, versus 30+ for question-end states). What is missing is transport: that identity must reach the generation position. Native heads do not reliably do it for novel entities, so the engine *constructs a head*.

A donor-head scan finds attention groups whose ablation does not hurt (often slightly improves) prose; the organ overwrites one at layer 24. Its QK are rank-one outer products: $W_q = a\, c\, \hat{\mu}^\top$ (queries fire from every position along a carrier direction c), $W_k = b\, c\, g^\top$ (g a subject-token direction, so keys are large only at subject mentions). The carrier's support on the rotary spectrum *shapes the kernel*: slow dimensions give a flat any-distance floor; a mid band (periods of roughly 60–250 tokens) adds constructive interference only at small relative distance — *recency selection*. At true conversation replays the head places 0.55–0.88 of its attention mass on the correct subject's tokens, resolving pronouns by discourse recency with zero additional parameters. Per-subject value lanes write orthonormal payload directions (chosen in the low-variance subspace of the residual, SNR ≈ 13 vs. random directions) that downstream readout gates consume one layer later — writing at layer 24 for layer-25 readers because payloads do *not* survive many blocks at readable strength (§7). Selector directions are orthogonalized across subjects; before that fix, shared person-name components caused cross-subject leakage.

3.7 Delivery insurance at the last layer

Some contexts fight back: for peak-prior continuations the native circuitry above layer 25 *amplifies against* an injected push — boosting the answer raised its competitor faster. The engine's answer is positional: a small rescue bank at layer 31, one RMSNorm from the logits, where no native layer remains to respond. Rescue neurons key on the exact failing context state, with generic-question and prose negatives, and carry orthogonalized answer-minus-competitor-centroid values sized by the measured gap. The same site hosts the *denial gag*: instruct models carry a reflex that answers unknown-person questions with "I couldn't find information on…"; one neuron keyed on the question-mode ramp subtracts the denial-opener direction, activation-normalized on delivered states (a mis-normalized first attempt over-suppressed 2.5× and produced attractor junk — the referee caught it).

3.8 Composition: hop enrichment

A fact you can only recite is weaker than a fact you can *use*. One further neuron per fact, placed at 0.16× depth (layer 5 on Llama-8B — the band where native subject enrichment lives), keys on the subject's state and adds the *answer entity's own representation* to the stream: after "*Marcus lives in Lyon*" is installed, the subject's tokens carry Lyon's representation, and the model's native machinery can answer "*what continent does Marcus live on?*" without the fact ever being stated. Attribution A/B on delivered checkpoints: relations never baked (continent, language of city of residence) go from 2/12 and 0/12 without enrichment to 11/12 and 11/12 with it — *above* the ∼6/10 rate at which a 6.9B model composes its own native facts silently. An ordering law governs the stage: enrichment must be written *before* every key harvest, because the engine must key on the model it is delivering — retrofitting it after harvests broke canonical recall entirely.

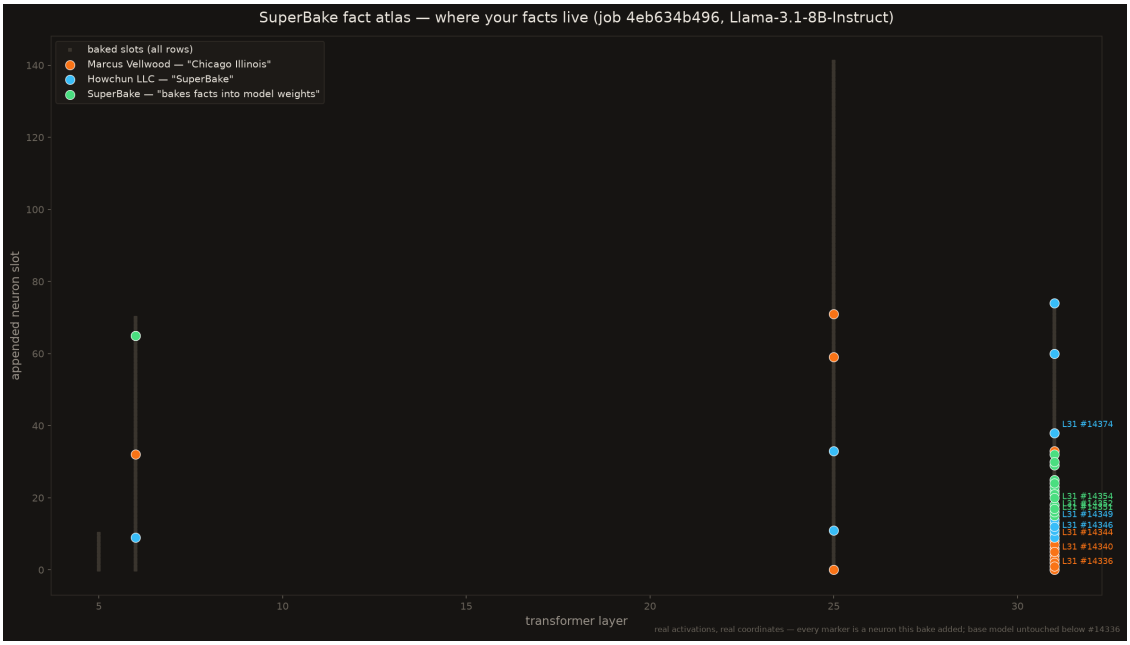

Figure 2: A receipt rendered as a map: every installed fact's neurons at their layer/slot coordinates in one real bake. The public demo at albertmi.ai lets visitors click any of 500 Wikipedia facts and fly to its neurons.

3.9 Verification, the pruner, and the receipt

Every bake ends in an examination the model must pass *as delivered*:

- **Battery**: generated phrasing variants per fact — primary, casing, alias, profile ("tell me about X"), and reverse (answer→subject) rows; teacher-forced checks are confirmed by greedy *generation*, which catches chain deaths that teacher-forcing hides.
- **Referees**: held-out known-fact questions (the model must still know what it knew), prose NLL on held-out text, and chat NLL on a held-out transcript — run after every engine stage with automatic revert of any stage that regresses them; plus a plain-recall spot check, added after a stage once passed every global referee while silently breaking per-row recall.
- **The pruner**: every row that fails verification is *neutralized* — key, bias, up row, and value column zeroed (shared code neurons only when all their readers failed) — and the battery re-runs. Failed facts degrade to the stock model's honest ignorance. A broken row left in the weights is worse than no row: it produces garbage near its key. *Failed* $\neq$ *absent* is a delivery invariant.
- **The receipt**: a JSON artifact listing every fact, per-category verification results, referee outcomes, and the layer/slot coordinates of every neuron of every fact (Fig. 2). Claims about borderline counts use best-of-N confirm runs; GPU nondeterminism alone moves borderline tallies by ±5–9 rows per 300.

4 Results

Against a strong trained baseline. The comparison target is not naive fine-tuning (which destroys the host model) but the best gradient-based system we could build: SGD confined by

Model	Battery	Verified	Prose NLL	Notes
Llama-3.1-8B-Inst. (SGD baseline)	7,450 rows	76.9%	2.97 → 7.24	masked, caged, guarded
Llama-3.1-8B-Inst. (constructed)	7,450 rows	**89.2–91.3%**	untouched	zero gradients
Llama-3.1-8B-Inst. (delivered ckpt)	300 + variants	93.3% / 89.5%	exact	primary / variants
Qwen2.5-7B-Inst. (delivered ckpt)	300 + variants	**97.3% / 95.5%**	exact	primary / variants
Qwen2.5-7B base (live queue)	65 rows	100%	+0.004	cleanest config
Pythia-6.9B (experimental tier)	300 + variants	77.3% / 75.0%	exact	§6

Table 1: Headline recall across models. "Delivered ckpt" rows are measured on saved, stock-reloaded checkpoints in a fresh process. The 7,450-row battery covers 1,000 facts with phrasing, casing, alias, profile, and reverse variants; the production engine run (89.2%) completed in 13.7 minutes including full verification and save on a single 80 GB-class GPU. Known-fact referees pass 4/4 and chat NLL *improves* on the constructed runs.

hard gradient masks to the same appended slots, with mixed batches, per-step guards, and an audit that reverts harmful writes. Table 1 and Fig. 3: construction wins overall (91.3% vs. 76.9%) and on every variant category, while the baseline's recall is bought with a 2.97 → 7.24 prose NLL collapse and construction's is free. The reversal-curse row deserves emphasis: models trained on "A is B" notoriously fail "B is A" [13]; written weights have no such asymmetry — the engine simply also writes the reverse rows, and they verify at ∼94%.

Durability and specificity. Under continued fine-tuning at sane learning rates, installed facts persist without protection; under deliberately abusive fine-tuning, constructed facts *outlive the model's own pretrained knowledge*. A 1,200-row twin-stress battery (300 subjects × 4 near-identical facts — the classic bleed geometry) verifies at 91.7% on Llama with *zero* measured twin-bleed (0/25 probes): whitened per-fact keys separate same-subject facts on relation words alone. Nonce installs work: a model that has never seen the token sequence "Zarvantis" learns to spell and use it (4/4).

Speed and scale. Construction is measurement-bound, not optimization-bound. The production engine bakes and fully verifies 1,000 facts in ∼14 minutes on one GPU. An earlier, reduced pipeline (recognition + naming neurons only, on raw Pythia-6.9B) wrote 192 facts in 13.0 seconds — 14.8 facts/second — with zero interference at that scale, and reached 2,000 facts at 92.0% verified in 139 seconds with known-fact referees clean and prose NLL slightly improved. On WikiBigEdit [8] real-world update questions, a 1,000-fact sample verifies at 859/1,000 — exceeding the 719 our best gradient-based system managed on the same sample at roughly 700× the compute.

The public artifact. The pipeline runs as a self-serve service (albertmi.ai): visitors submit facts, watch the bake, chat with the result, and download the checkpoint with its receipt from Hugging Face. A 500-fact Wikipedia atlas (Fig. 2) serves as a standing, inspectable demo bake.

5 What SGD Actually Builds

SuperBake's design was not guessed; it was *copied* from a mechanism we first had to discover. When the strong SGD baseline crams ∼1,000 facts into appended slots, the entire weight delta is observable: we own the seed state, the gradient mask confines every update to known coordinates, and the post state can be dumped and dissected. The post-mortem overturned our assumptions three times, and each reversal is load-bearing for the final method.

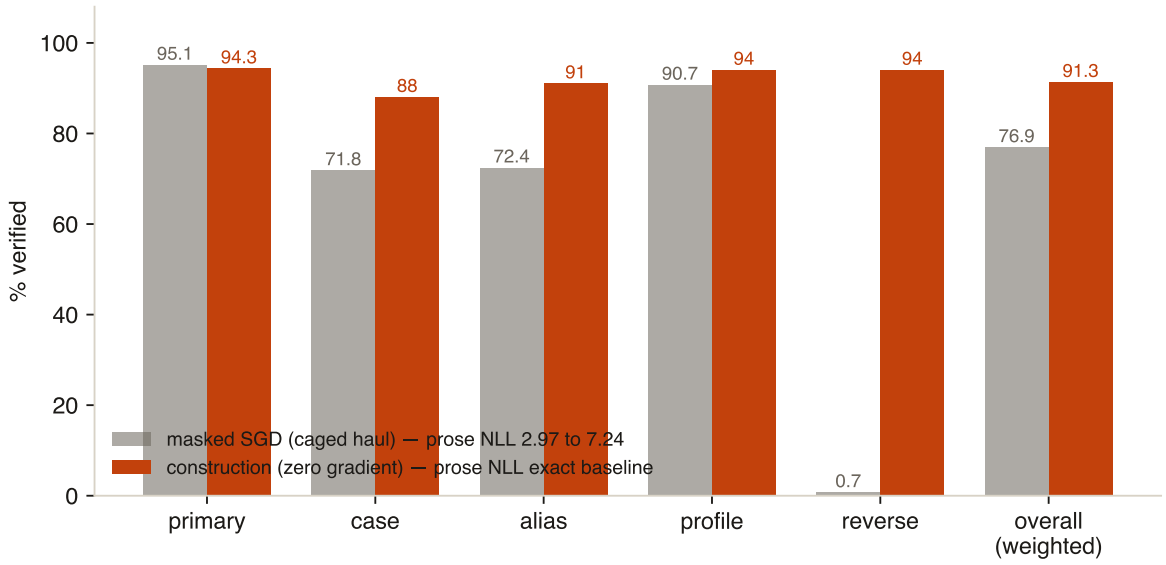

Figure 3: Per-category verification, Llama-3.1-8B-Instruct, 7,450-row battery. Grey: the strong baseline — SGD confined by gradient masks to appended slots, guarded and audited (its unmasked variant destroys the model outright). Orange: full construction. The baseline pays 2.4× prose NLL for its recall; construction pays nothing. Reverse questions collapse under training (0.7%) and are trivial under construction (∼94%): the reverse mapping is simply written as its own rows.

No neuron contains its fact. Across 3,787 trained neurons, the trained value column's top pushed token is the fact's own answer in under 0.3% of cases; per-neuron statistics of neurons whose facts came alive versus stayed dead are *indistinguishable*. Storage is collective. Pools compose super-additively (early pool alone +17.9 answer logits, late pool alone +4.3, together +30.4), and the early deposit acts as a *selectivity field*: with it, exactly one of 207 candidate late neurons fires per question; without it, a cross-fire soup.

The facts live in the leak. The decisive experiment (Fig. 4): sharpening every trained gate into a hard threshold — a transformation that leaves supra-threshold behavior untouched and only removes silu's sub-threshold leak — fully heals prose (6.80 → 2.69 NLL, stock = 2.64) and collapses verification from 270/300 to 3/300. Dense fact storage under SGD is a *coherent analog population code carried by activation leakage*: thousands of neurons each responding below threshold, their coherent sum encoding the answers. This explains, at once, why no column reads out an answer, why alive and dead neurons look identical, why removing the single shared "answer-mode" direction from all columns repairs prose *and* kills half the facts (one direction is simultaneously the prose poison and the storage medium), and why the prose tax grows with stored volume — the medium operates on prose too. It also sets a capacity law: allowed to hum, the analog code holds ∼9× more facts than the same slots restricted to firing neurons under key crowding.

Transport laws. Hand-written cross-layer signals fail in specific, measurable ways: directions chosen in *quiet* subspaces (to be readable over natural traffic) are attenuated ∼30× by the intervening network; *loud* directions travel with gain 1.9 but arrive rotated into shared, crowded channels. SGD's advantage is co-training both ends through the real Jacobian. The constructed answer is to stop fighting the medium: make separability *manufactured* (injected codes, §3.3) rather than found, and read every payload one layer after it is written (§3.6).

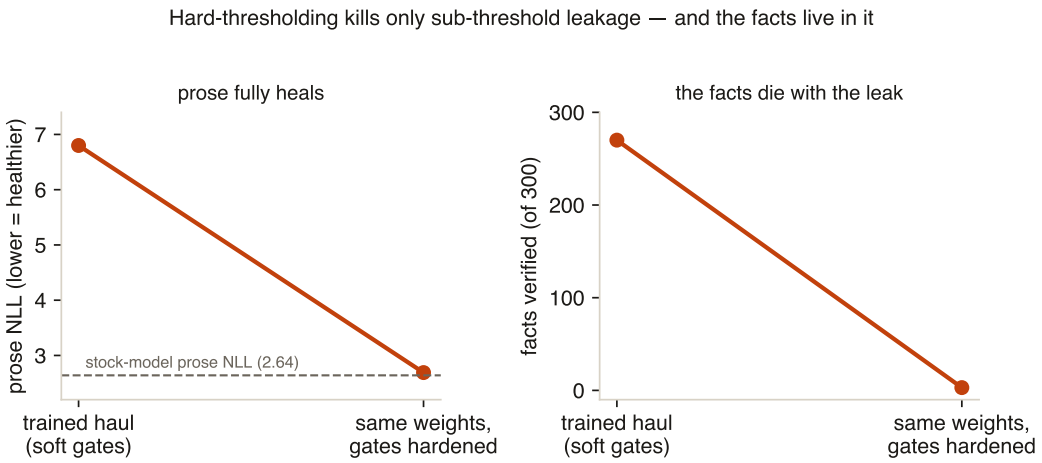

Figure 4: The leak-substrate experiment (Llama-8B, 1,000-row trained checkpoint). Hardening every trained gate — which changes *nothing* above threshold and only kills sub-threshold leakage — returns prose to the stock baseline and simultaneously destroys 267 of 270 verified facts. The facts were living *in* the leak.

The derived rule, then the ladder. Functionally, the trained solution is a *mode-gated linear associative memory* operating in the leak zone — and Eq. 1 shows the MLP can express exactly that object *gated*: gate = question-mode key, up = key row, value = value column; summed, mode $\cdot$ (VKx), a Kohonen heteroassociator [12] with a closed-form solve. The first constructed replacement was that literal solve. It plateaued at $\sim$58% — a single-layer linear map fights with only the layers above it, while SGD's early deposits ride nineteen native nonlinear layers of amplification; *the network is the kernel, and it is upstream.* Capacity comes from many *thresholded keys*, not matrix rank. Rebuilding along those two lines — per-fact hard-keyed neurons at both ends plus the closed loop of §3.5 — produced the ladder of Fig. 5: ten design rungs from 57.7% to 96.2%, past the SGD reference, at exactly zero prose cost. Every rung was forced by a measured failure, and the five constructed attempts that *preceded* the discovery of the leak substrate all failed for the same reason in different costumes: they wrote digital circuits into what is natively an analog medium.

6 Cross-Architecture Portability

The recipe is stated in terms any decoder transformer exposes — states, unembeddings, MLP slots — but each family taught a law.

Qwen2.5 (no MLP biases). Without biases there is nowhere to put a threshold. Folding thresholds onto the residual stream's constant carrier direction fails at exactly the positions that matter: the carrier is a *phantom at the tail* (attention-sink and first-token positions carry $\approx$0 carrier projection, evaporating any folded threshold). The fix is structural: project keys perpendicular to the carrier and let the *quadratic* gate$\times$up form supply selectivity — own-fact states score $\sim$6.5 against prose ≤ 0.3, and the square makes the leak ratio $(0.3/6.5)^2 \approx 0.2\%$. Threshold-free, bias-free, and the delivered checkpoint is the best of all models (97.3%/95.5%).

Pythia (GPT-NeoX, thin margins). Pythia exposed the sharpest law (Fig. 6): *born-hard banks amplify composition wobble.* Cross-context state jitter of $\pm$0.2 units is harmless at Llama's margins ($\sim$7) and catastrophic at Pythia's ($\sim$0.9–2.6): a flipped near-threshold neuron is a flipped

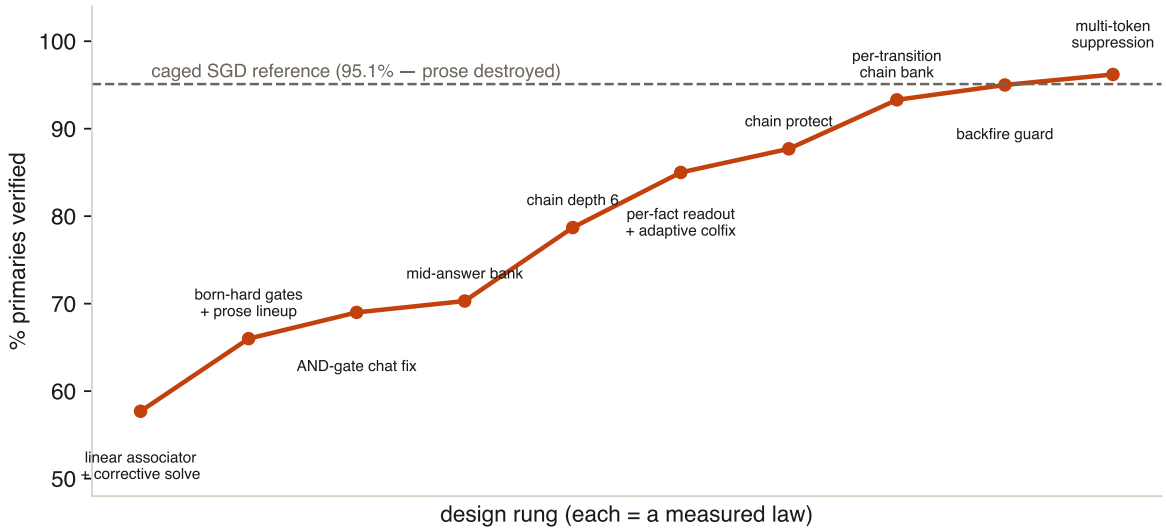

Figure 5: The construction ladder (Llama-3.1-8B, 300 primary facts, chat battery). Each rung is a design change forced by a measured failure mode of the previous rung. Throughout the climb three invariants hold: zero gradient steps, prose NLL at exact stock baseline, and end-to-end build time under ∼2.5 minutes.

8–28 logit push, so a calibration loop that measures in one batch composition converges to a mirage — gaps closed in its own lighting, absent in truth. The cure is the multi-lighting harvest of §3 taken to its limit: keys on the mean across lightings, own-thresholds at the minimum, negatives at the maximum, and every repair round measured under a *fresh* shuffle. Pythia climbs 7 → 77.3% — above the SGD engine's Llama number, on a harder skeleton — and is shipped as an experimental tier.

7 Design Laws

Each law in Table 2 was paid for with a refuted design; together they are the field manual this paper exists to publish.

8 Limitations and the Honest Boundary

Free conversation on bare weights. The strongest honest claim is scoped: installed facts answer *any baked phrasing*, including mid-conversation (96–98% transcript-render recall), and paraphrase generalizes. But fully free recall — pronouns, ellipsis, novel phrasings deep in a dialogue whose history the model itself generated — is measurably harder: on a 20-turn adversarial no-crutch battery the delivered weights answer 5–8/20, and four independent non-transport interventions each cap at that band. The measured causes are distributional (self-generated conversational states are not covered by any scripted harvest — the bench/battery divergence is itself the measurement) and mechanical (delivery keys go key-dead at depth; the attention organ of §3.6 is the constructed response and has produced the first depth hits, e.g. pronoun follow-ups, while raising in-battery verification to its all-time high). Delivered models disclose `verified_conversational` per fact in the receipt, and ship with a small client-side phrasing-rescue script; the serving layer never answers from the receipt.

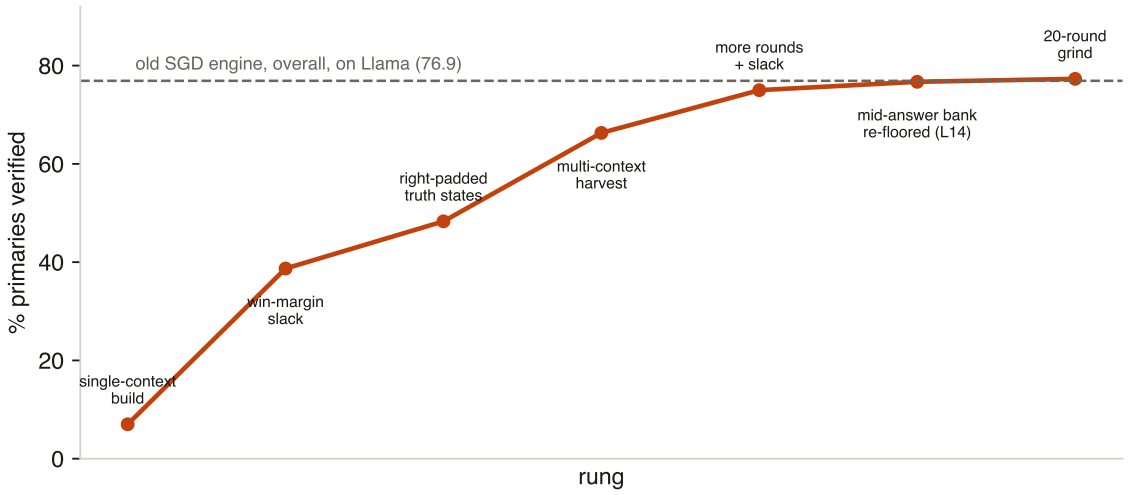

Figure 6: Pythia-6.9B and the context-binding law. A build whose keys are harvested in one batch composition verifies at 7% in any other context — the pushes exist only under the training lighting. Each rung generalizes the harvest (multiple lightings, worst-case thresholds, verification under fresh compositions) until the constructed facts win in single-sequence reality.

Other boundaries. Answers that begin with the subject's own name stress chain keys (an engine fix is queued); one subject cannot carry two facts phrased as the same question (the augmenter now deduplicates); Qwen carries a +1.6-nat chat-template NLL gap (diffuse, cosmetic, compensator designed); Pythia's tier trails the production families; and the largest single bake certified here is 1,000 facts (2,000 on the reduced pipeline) — the behavior of this engine at 10^5 facts is unmeasured. All numbers are from a single-author engineering campaign: independent replication is the point of releasing the method.

9 Related Work

Knowledge editing. ROME and MEMIT [2, 3] edit existing MLP projections via rank-constrained least squares; MEND [4] and KE [5] learn editor networks; [6] fine-tunes with constraints. SuperBake differs in substrate (appended, initially-silent slots; originals untouched), in mechanism (closed-form construction plus gradient-free calibration), in contract (behavioral verification with neutralization and receipts), and in scope (multi-token answers, variants, reverse rows, conversation renders, composition). WikiBigEdit [8] finds editing methods underperforming retrieval and fine-tuning at lifelong scale — consistent with our reading that editing existing weights competes for shared capacity, which appending avoids. **MLPs as memories.** Our construction is a literal engineering of the key–value view of transformer MLPs [1] and of classical associative memories [12]; knowledge-neuron attribution [7] similarly localizes facts in MLPs. **Mechanistic interpretability.** The circuit formalism [9], logit-lens readouts [11], and superposition [10] supply the vocabulary; our leak substrate (§5) is a superposed *analog* code, and our instruments are standard MI practice turned into engineering referees. **Alternatives.** RAG [14] leaves weights unchanged; LoRA [15] trains low-rank deltas with SGD's interference profile. The reversal curse [13] motivates our reverse rows. Model families: Llama 3 [18], Qwen2.5 [19], Pythia [17], RoPE [16].

Law	Statement
Leak substrate	Dense trained fact storage is a coherent sub-threshold population code in activation leakage; prose damage and storage are the same signal (Fig. 4).
Capacity	Recall capacity scales with the number of *thresholded keys*, not matrix rank; shared linear maps saturate near the hidden dimension.
Transport	Quiet directions attenuate $\sim$30$\times$; loud ones arrive rotated into shared channels. Manufacture separability (codes); never rely on found channels.
Payload survival	Read a constructed payload at the layer where it is written +1; it does not survive many blocks at readable strength.
Whitening	Matched-filter keys need union covariance (facts + innocents); raw mean keys measure the shared question subspace.
Born-hard	Every hand-set gate must be hardened ($\times s$, $/s$); protection lives in gates with clearance floors, never in solve-side zero targets.
Per-row keys	Group-mean keys are diluted: thresholded at the worst phrasing, oversized for the rest. One row, one key.
Sibling sum	Same-transition siblings sum at generation; each carries $1/n$ of the push.
Ordering	Later writes shift earlier measurements: enrichment before harvests; readout repair before chain repair; earliest failing transition first.
Context binding	Keys harvested under one context distribution exist only there when margins are thinner than state wobble; harvest under many lightings, threshold at worst case, verify under fresh ones.
RoPE locality	Rank-one rotary QK kernels peak only at distance zero and oscillate elsewhere: constructed heads cannot implement local windows; content selectivity must carry them.
Referee completeness	Global referees (knowns, prose, chat) are blind to per-row damage; every stage needs a plain-recall spot check.
Delivery honesty	Failed $\neq$ absent: neutralize unverified rows; never let a serving layer answer for the weights.
Confirm runs	GPU nondeterminism moves borderline tallies $\pm$5–9 rows per 300; single-run records are noise until confirmed.

Table 2: Design laws for weight-space fact engineering, each derived from a measured failure documented in the engineering ledger.

10 Conclusion

Facts can be written into a transformer the way an engineer writes to memory: measured, addressed, verified, and reversible. The write head is not gradient descent — it is a set of closed-form constructions that speak the network's own mechanism, discovered by dissecting what gradient descent does when asked to cram. The resulting system beats its trained counterpart on recall while refusing the damage trained methods accept as inevitable, dissolves the reversal curse by construction, and turns "what does the model know?" from a probing question into a lookup in a receipt. The open frontier is conversational: making installed knowledge as *reachable* in free dialogue as native knowledge is. The instruments, laws, and the attention-organ line of attack published here are, we hope, enough for others to close it with us.

Acknowledgments. The engineering campaign behind this paper was conducted with substantial assistance from Claude (Anthropic) as a research and engineering agent.

References

[1] M. Geva, R. Schuster, J. Berant, O. Levy. Transformer Feed-Forward Layers Are Key-Value Memories. *EMNLP*, 2021.

[2] K. Meng, D. Bau, A. Andonian, Y. Belinkov. Locating and Editing Factual Associations in GPT. *NeurIPS*, 2022.

[3] K. Meng, A. S. Sharma, A. Andonian, Y. Belinkov, D. Bau. Mass-Editing Memory in a Transformer. *ICLR*, 2023.

[4] E. Mitchell, C. Lin, A. Bosselut, C. Finn, C. D. Manning. Fast Model Editing at Scale. *ICLR*, 2022.

[5] N. De Cao, W. Aziz, I. Titov. Editing Factual Knowledge in Language Models. *EMNLP*, 2021.

[6] C. Zhu et al. Modifying Memories in Transformer Models. *arXiv:2012.00363*, 2020.

[7] D. Dai, L. Dong, Y. Hao, Z. Sui, B. Chang, F. Wei. Knowledge Neurons in Pretrained Transformers. *ACL*, 2022.

[8] L. Thede, K. Roth, M. Bethge, Z. Akata, T. Hartvigsen. WikiBigEdit: Understanding the Limits of Lifelong Knowledge Editing in LLMs. *ICML*, 2025. arXiv:2503.05683.

[9] N. Elhage et al. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*, 2021.

[10] N. Elhage et al. Toy Models of Superposition. *Transformer Circuits Thread*, 2022.

[11] nostalgebraist. Interpreting GPT: the Logit Lens. *LessWrong*, 2020.

[12] T. Kohonen. Correlation Matrix Memories. *IEEE Transactions on Computers*, C-21(4), 1972.

[13] L. Berglund et al. The Reversal Curse: LLMs Trained on "A is B" Fail to Learn "B is A". *arXiv:2309.12288*, 2023.

[14] P. Lewis et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *NeurIPS*, 2020.

[15] E. Hu et al. LoRA: Low-Rank Adaptation of Large Language Models. *ICLR*, 2022.

[16] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, Y. Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing*, 2024.

[17] S. Biderman et al. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. *ICML*, 2023.

[18] A. Grattafiori et al. The Llama 3 Herd of Models. *arXiv:2407.21783*, 2024.

[19] Qwen Team. Qwen2.5 Technical Report. *arXiv:2412.15115*, 2024.