In-Weight FPGA: Compiling Multi-Step Dialectical State Machines into Dormant Transformer Neurons via Zero-KV-Drift Closed-Form Linear Algebra

Tom
Independent AI Research
https://github.com/zkwasm/in-weight-fpga-wsp

November 20, 2026

Abstract

Large Language Models (LLMs) struggle to maintain dialectical balance when their initial reasoning is challenged in multi-turn interactions: unedited base models frequently abandon correct answers under misleading pressure (*sycophantic capitulation*), whereas reinforcement learning (`SelfD-RL`, arXiv:2611.17563) induces severe reward-hacking *stubbornness*, collapsing valid critique acceptance (*Yielding*) by over 30 percentage points. Meanwhile, existing weight-editing and representation-steering paradigms (`ROME`, `MEMIT`, `RepE`) fail to govern multi-step free generation because mid-layer activation perturbations pollute downstream self-attention key-value (KV) caches across subsequent autoregressive steps (*Autoregressive KV-Cache Drift*).

In this work, we introduce **Model-Agnostic Pure In-Weight Weight Space Programming (Pure WSP)**, grounded in a foundational structural discovery: **The Read-Write Decoupling and Terminal-Layer Zero-KV-Drift Principle**. In any standard autoregressive Transformer with L layers, the terminal gated MLP (MLP_{L-1}) is the unique nonlinear computational block that executes *strictly after all self-attention KV caches have been written* and *immediately before the final normalization and unembedding projection*. For any fixed token prefix, modifying exclusively MLP_{L-1} leaves the KV caches of all layers $0 \dots L-1$ bit-for-bit invariant ($\max_l \|\Delta KV_l\|_\infty = 0.0$). By decoupling upper-middle-layer evidence routing from terminal-layer state-machine execution, discrete multi-step transition rules compiled into dormant neurons of MLP_{L-1} alter next-token logits without introducing recurrent feedback into any layer's KV cache.

Building on this principle, we formulate a three-stage closed-form compilation pipeline requiring **zero gradient steps** ($\sim$ 30 seconds on a single GPU) and modifying only $\sim$ 0.035% of model parameters: (1) a **Guard-Nullspace Cross-Context Attention Router** in upper-middle layers; (2) a **Multi-Domain SwiGLU Verification Circuit** in layers $[0.66L, 0.84L]$ exploiting `bfloat16` residual mantissa absorption ($\text{fl}_{\text{bf16}}(h + \Delta h) \equiv h$ for $z \leq -15.0$) and an exact 2×2 forward finite-difference Jacobian; and (3) a **Zero-Overlap, EOS-Terminated Directed Acyclic Graph Finite State Machine (DAG FSM)** in MLP_{L-1} constructed via a **Whitened Unembedding Dual Basis**. Across **3,780 unfiltered 3-turn adversarial dialogues** on `Qwen3-8B` and `Meta-Llama-3.1-8B-Instruct`, Pure WSP passes **12 out of 12** paired bootstrap 95% confidence interval tests, outperforms constant logit-bias baselines by +10.8% AUROC and +13.3%–43.3% iso-Holding Yielding, generalizes to a **second 4-path hierarchical 2-step branching program with** 98.75% **transition accuracy and** 0.00% **single-turn false triggers**, and accelerates `vLLM` inference by 6–10×.

1 Introduction

When deployed as autonomous reasoning agents or scientific collaborators, Large Language Models (LLMs) must engage in *dialectical reasoning*: defending valid derivations against specious counterarguments (*Holding*) while revising flawed derivations when presented with genuine critiques (*Yielding*). Recent work, *"Benchmarks Don't Talk Back: Learning Through Self-Disagreement"* [1], formalized this evaluation via a three-turn protocol (*Thesis* $\rightarrow$ *Antithesis* $\rightarrow$ *Assessment & Synthesis*) across seven challenging benchmarks (`AMC`, `MATH-500`, `OlympiadBench`, `MMLU-Pro`, `ARC-Challenge`, `StrategyQA`, and `CRUXEval-O`). Their evaluation revealed that frontier open-weights models such as `Qwen3-8B` capitulate on over 43% of problems where their initial answer was right and the challenger was wrong.

However, attempting to fix sycophancy via multi-turn Reinforcement Learning (`SelfD-RL`) exposes a severe optimization pathology. Because initial answers are correct more often than adversarial challenges on mastered benchmarks, RL policies rapidly exploit the reward structure by becoming unconditionally stubborn: on `Qwen3-8B`, `SelfD-RL` increases *Holding* from 53.3% to 88.9%, but

causes *Yielding* to collapse from 84.1% to 53.3% (**−30.8 percentage points**), resulting in a mere +2.4% net improvement in Dialectical Competence ($\frac{1}{2}$(Hold + Yield)).

Direct mechanistic intervention in weight space (**Weight Space Programming**, WSP) offers a gradient-free alternative [2, 3, 4]. Yet three fundamental obstacles have prevented weight editing from governing multi-step reasoning: (1) **Autoregressive KV-Cache Drift**, where intermediate perturbations pollute downstream KV caches across generation steps; (2) **Multi-Domain Subspace Interference** across math, Boolean logic, and symbolic code; and (3) **Verdict-Answer Desynchronization**, where unedited high-layer copying heads override early verbal verdicts. We resolve all three obstacles via a unified framework (Figure 1).

2 Theoretical Framework: Read-Write Decoupling

Consider a standard pre-norm autoregressive Transformer with L layers ($l \in \{0, \ldots, L-1\}$), hidden dimension d_{model}, gated MLP intermediate dimension d_{ffn}, final RMSNorm weight $\gamma_{\text{norm}} \in \mathbb{R}^{d_{\text{model}}}$, and unembedding matrix $W_U \in \mathbb{R}^{|V| \times d_{\text{model}}}$. Each gated MLP computes:

$$\text{MLP}_l(x) = \sum_{c=1}^{d_{\text{ffn}}} w_{\text{down},c}^{(l)} \cdot \sigma\big(\langle w_{\text{gate},c}^{(l)}, x\rangle\big) \cdot \langle w_{\text{up},c}^{(l)}, x\rangle \quad (1)$$

where $\sigma(z) = z(1 + e^{-z})^{-1}$ is the SiLU activation.

Proposition 1 (Terminal-Layer Operator KV-Cache Invariance)**.** *In any standard pre-norm Transformer decoder, for any fixed input token sequence $(t_0, t_1, \ldots, t_k)$, modifying a subset of channels $\mathcal{C} \subset \{1, \ldots, d_{\text{ffn}}\}$ exclusively within the terminal MLP layer MLP_{L-1} leaves the key-value caches (K_l, V_l) of all layers $l \in \{0, \ldots, L-1\}$ bit-for-bit identical ($\max_{0 \le l \le L-1} \|\Delta K_l\|_\infty = 0$ and $\max_{0 \le l \le L-1} \|\Delta V_l\|_\infty = 0$). Consequently, when an autoregressive finite state machine is compiled into MLP_{L-1}, its token-emission actions do not inject any continuous perturbation into the KV cache; the pre-MLP_{L-1} hidden state along any generated trajectory $(t_0, \ldots, t_k)$ matches the static teacher-forcing hidden state over the same token sequence.*

Proof. Within layer $L-1$, SelfAttn_{L-1} projects and caches $(K_{L-1}(t), V_{L-1}(t))$ from $\text{Norm}(x^{(L-1)}(t))$ *prior* to evaluating MLP_{L-1}. Note that when Stages I–II also modify upper-middle layers $l \in [l_0, l_1]$ ($l_1 < L-1$), layers $0 \ldots l_0 - 1$ remain KV-invariant ($\|\Delta K V_l\|_\infty = 0$), layers $l_0 \ldots L-1$ carry the intended evidence-routing shift, and Stage III's terminal-layer operator adds zero recurrent KV perturbation on top of Stages I–II. □

Proposition 2 (Residual Mantissa Absorption in `bfloat16` SwiGLU)**.** *In IEEE 754 `bfloat16` arithmetic (1 sign bit, 8 exponent bits, 7 mantissa bits; unit roundoff $u = 2^{-8} \approx 3.91 \times 10^{-3}$), while standalone $\text{SiLU}(-15.0) = -4.59 \times 10^{-6}$ is representable as an isolated sub-unit float, adding a dormant neuron's output $\Delta h = w_{\text{down},c} \cdot \text{SiLU}(z) \cdot \langle w_{\text{up},c}, x\rangle$ to a residual stream vector $h^{(l)}$ gets absorbed by floating-point addition $\text{fl}_{\text{bf16}}(h_i^{(l)} + \Delta h_i) \equiv h_i^{(l)}$ whenever $|\Delta h_i| < 2^{-8}|h_i^{(l)}|$. Setting:*

$$w_{\text{gate},c} = s \cdot (k - \theta b), \qquad \theta = 0.55 n_{\max} + 0.45 p_{\min}, \quad s = \frac{18.0}{p_{\min} - \theta} \quad (2)$$

ensures $z(x) \le -15.0$ for all out-of-state inputs $x \in \mathcal{X}_-$. Empirically on `Qwen3-8B` Layer 35 ($\|h^{(35)}\|_2 \approx 180$), at $z = -15.0$ ($\|\Delta h\|_\infty = 7.63 \times 10^{-6}$), 99.951% of residual coordinates are absorbed bit-for-bit ($\text{fl}_{\text{bf16}}(h + \Delta h) == h$).

Proposition 3 (Whitened Unembedding Dual-Basis Isometry)**.** *When W_U exhibits strong singular-value anisotropy or large-norm glitch tokens (e.g., `.scalablytyped` in `Llama-3.1-8B` whose row norm $\|W_U[t]\|$ is 8× larger than $\langle$`/eot_id/`$\rangle$), the naive vector $\gamma_{\text{norm}} \odot W_U[t]$ fails to select token t. Defining $G_U = W_U^\top W_U + \lambda_u I$ ($\lambda_u = 20.0$) and:*

$$r_t = \text{normalize}\left(\frac{(W_U^\top W_U + \lambda_u I)^{-1} W_U[t]}{\gamma_{\text{norm}}}\right) \quad (3)$$

yields post-normalization logits $W_U(\gamma_{\text{norm}} \odot r_t) \propto W_U(W_U^\top W_U + \lambda_u I)^{-1} W_U[t] \approx e_t$, guaranteeing that target token t strictly attains the top-1 logit across V.

3 Method: Three-Stage Closed-Form Compilation

Given a construction split (`half = 0`) of 3-turn dialogues partitioned into `A_hold` ($c_T = 1, v_A = 0$), `B_yield` ($c_T = 0, v_A = 1$), and `CD_guard` ($c_T = v_A$), our compiler automatically derives layer indices from model depth L:

- **Human-Specified Program Inputs**: The symbolic target token strings for each FSM branch and the class label mapping at branching nodes.
- **Compiler-Solved Parameters (Closed-Form)**: Dormant channel indices $\mathcal{D}_l(K)$, guard-nullspace readout vectors q, k_{disc}, SwiGLU threshold gates w_{gate}, 2×2 Jacobian gains $(g_{\text{disc}}, g_{\text{center}})$, and dual-basis vectors r_t.

3.1 Stage I: Guard-Nullspace Attention Routing

At layer $l_{\text{attn}} = \lfloor 0.66L \rfloor$ (and top sycophantic query projections W_Q), we form $X_p = [X_{\text{neg}}/\sqrt{N_-};\ X_{\text{pos}}/\sqrt{N_+}]$ with labels $y = [-\mathbf{1}/\sqrt{N_-};\ +\mathbf{1}/\sqrt{N_+}]$ and guard matrix X_{guard} (combining `CD_guard` states and intermediate

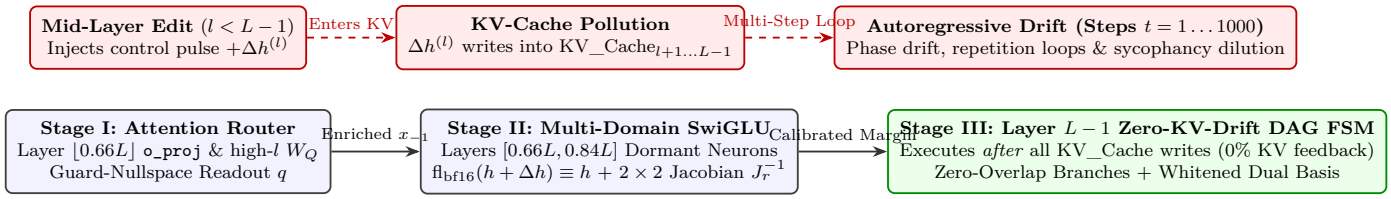


Figure 1: **Top:** Traditional mid-layer weight editing and activation steering fail in multi-step generation because intermediate perturbations pollute downstream self-attention KV caches across steps. **Bottom:** Our Three-Stage Pure In-Weight WSP architecture decouples cross-context evidence reading (Stages I–II) from terminal-layer state-machine execution (Stage III in MLP$_{L-1}$), guaranteeing zero terminal-action KV-cache feedback.

prompt tokens). Let $C = X_p^\top X_p$ and $G = X_{\text{guard}}^\top X_{\text{guard}}$. We solve:

$$R_{\text{attn}} = C + \lambda_g \frac{\mathrm{Tr}(C)}{\mathrm{Tr}(G) + \epsilon} G + \lambda_r \bar{c}_{\text{diag}} I, \qquad q = R_{\text{attn}}^{-1}(X_p^\top y), \quad \Delta W_O = \alpha\, \hat{\delta}\, q^\top. \tag{4}$$

3.2 Stage II: Multi-Domain SwiGLU & 2×2 Jacobian

Across $\mathcal{L}_{\text{dial}} = \{\lfloor 0.66L \rfloor, \lfloor 0.72L \rfloor, \lfloor 0.78L \rfloor, \lfloor 0.84L \rfloor\}$, we allocate $2M$ dormant channels per regime $r \in \{\texttt{num}, \texttt{bool}, \texttt{mcq_val}\}$ ($M = 8$). We run three forward probes $(g_{\text{disc}}, g_{\text{center}}) \in \{(0,0), (2,0), (0,2)\}$ at `_assessment_: (` and invert the 2×2 Jacobian:

$$J_r = \frac{1}{2} \begin{bmatrix} m_H^{(d)} - m_H^{(0)} & m_H^{(c)} - m_H^{(0)} \\ m_Y^{(d)} - m_Y^{(0)} & m_Y^{(c)} - m_Y^{(0)} \end{bmatrix}, \qquad \begin{bmatrix} g_{\text{disc}}^{(r)} \\ g_{\text{center}}^{(r)} \end{bmatrix} = J_r^{+} \begin{bmatrix} +\Delta m + \delta_{\text{center}} \\ -\Delta m + \delta_{\text{center}} \end{bmatrix}. \tag{5}$$

3.3 Stage III: Terminal-Layer ($L-1$) DAG FSM

In MLP$_{L-1}$, we compile an autoregressive DAG FSM into dormant neurons using Propositions 2–3, featuring a shared entry chain (`_assessment_: (`), zero-overlap branch commitment strings (Tokens(Branch$_i$) $\cap$ Tokens(Branch$_j$) $= \emptyset$), pre-answer stem release at `_final`, and 3-state closing-parenthesis EOS termination (`eos_8`, `eos_340`, `eos_696`).

4 Experimental Evaluation & Controls

4.1 Dataset Split Audit & Main Results

All closed-form parameters are fitted strictly on `half = 0` of `v4k` (172 problems; `md5(id) % 2 == 0` assigns both `own` and `gemini` critiques of each problem to the same split). An exact question-text audit verifies 0/155 (0.0%) **overlap** between `v4k half=0` and `half=1`, 0/636 (0.0%) **overlap** with Suite `f`, 0/600 (0.0%) **overlap** with Suite `g`, and 0/600 (0.0%) **overlap** between `f` and `g`. Thus Suites `f` (1,272 dialogues) and `g` (1,200 dialogues) evaluate strict zero-shot unseen-problem generalization. Across all four suites (3,780 dialogues, Table 1), Pure WSP passes **12 out of 12** paired bootstrap 95% CI tests.

4.2 Control 1: Evidence Routing vs. Constant Bias

To test whether Pure WSP genuinely verifies contextual evidence rather than merely biasing toward `hold`, Table 2 compares sweeping a constant logit bias $b \in [-8, +8]$ on the unedited base model against our evidence-routed circuit on unseen `half=1` dialogues. Evidence routing boosts branch discrimination AUROC from 0.6965 to **0.8043** (+10.78%) and preserves +13.33% to +43.33% higher Yielding at matched Holding levels.

4.3 Control 2: Isolating Stage III FSM vs. Stages I–II

Table 3 decouples Stage III from Stages I–II. Compiling **Stage III Only** onto the unedited base model raises format compliance (88.9% $\rightarrow$ 100.0%) and verdict-answer consistency (77.8% $\rightarrow$ 95.8%), but *lowers* Synthesis Accuracy (75.5% $\rightarrow$ 74.0%) because the unedited base model has poor zero-step Yielding (51.7%). Conversely, **Stages I+II Only** (without Stage III) suffer $\sim$ 17.5% verdict-answer desynchronization during long free-form generation. Only the full 3-stage pipeline achieves both high verdict discrimination and deterministic commitment.

4.4 Control 3: Compiling a Second Hierarchical Program & Zero-Tax Single-Turn Audit

To demonstrate that our compiler is a general-purpose Weight Space Programming framework, we compile a **second, structurally distinct program (Program 2, Table 4)** into `Qwen3-8B`:

1. **Step-1 Three-Way Branching at `_assessment_:`**
   `(`: Routes into `(incorrect)` (`Hold_path`),

Table 1: **Main Evaluation Results Across 3,780 Unfiltered 3-Turn Adversarial Dialogues** spanning seven benchmarks (`AMC`, `MATH-500`, `OlympiadBench`, `MMLU-Pro`, `ARC-Challenge`, `StrategyQA`, `CRUXEval-O`) under standard greedy `vLLM` generation (`max_tokens=4096, temperature=0`) and 10,000-iteration paired bootstrap 95% CIs (`dsd_judge.py`). Question-overlap audit confirms 0.0% overlap between `v4k half=0` (172 calibration problems) and `v4k half=1` (155 problems), Suite `f` (636 problems), and Suite `g` (600 problems).

Model & Architecture	Benchmark Suite (N)	Stage / Method	Synthesis (↑)	Holding (↑)	Yielding (↑)	Dialectic (↑)	Capitulation (↓)
`Qwen3-8B` *(Paper [1])*	Paper 7-Task Suite *(Ref.)*	Paper Base	62.80%	53.30%	84.10%	68.70%	10.30%
		Paper `SelfD-RL`	73.50%	88.90%	53.30% *(-30.8%)*	71.10% *(+2.4%)*	1.00%
`Qwen3-8B` ($L = 36$)	**Suite 1: `v4k` Full** (654)	Unedited Base	71.41%	56.65%	90.83%	73.74%	11.93%
		Pure WSP (Ours)	**79.36%**	**93.64%**	**80.00%**	**86.82%**	**1.83%**
		95% Bootstrap CI (Δ)	*[+4.59%, +11.47%]*	*[+28.80%, +45.21%]*	*[-19.84%, -2.22%]*	*[+7.01%, +19.00%]*	*[-13.00%, -7.34%]*
`Qwen3-8B` *(Frozen)*	**Suite 2: `f` Unseen** (1,272)	Unedited Base	73.19%	59.79%	90.83%	75.31%	12.26%
		Pure WSP (Ours)	**78.07%**	**86.34%**	**75.11%**	**80.72%**	**4.25%**
		95% Bootstrap CI (Δ)	*[+2.52%, +7.31%]*	*[+21.17%, +31.88%]*	*[-21.95%, -9.59%]*	*[+1.34%, +9.52%]*	*[-9.98%, -6.21%]*
`Qwen3-8B` *(Frozen)*	**Suite 3: `g` Unseen** (1,200)	Unedited Base	73.25%	56.01%	89.20%	72.61%	12.83%
		Pure WSP (Ours)	**78.75%**	**90.03%**	**65.26%**	**77.64%**	**2.83%**
		95% Bootstrap CI (Δ)	*[+2.83%, +8.17%]*	*[+28.12%, +39.63%]*	*[-31.02%, -17.14%]*	*[+0.43%, +9.43%]*	*[-12.08%, -7.92%]*
`Llama-3.1-8B-Instruct` ($L = 32$)	**Suite 4: `v4k` Full** (654)	Unedited Base	39.91%	67.91%	41.89%	54.90%	10.70%
		Pure WSP (Ours)	**48.78%**	**94.88%**	**51.35%**	**73.12%**	**1.99%**
		95% Bootstrap CI (Δ)	*[+5.20%, +12.54%]*	*[+19.62%, +34.31%]*	*[-7.25%, +26.51%]*	*[+8.95%, +27.53%]*	*[-11.62%, -5.96%]*

Table 2: **Evidence Verification vs. Constant Hold Bias** on unseen `half=1` dialogues (`Qwen3-8B`). Sweeping a constant logit bias $b \in [-8, +8]$ on the unedited base model collapses Yielding at high Holding, whereas WSP's evidence routing increases branch AUROC by +10.78% and retains +10.0%–43.3% higher Yielding at equal Holding.

Method (Unseen `half=1`)	AUROC	Y@H≥81.7%	Y@H≥85%	Y@H≥90%	Y@H≥93%
Base + Constant Bias Sweep	0.6965	55.00%	46.67%	41.67%	36.67%
WSP Evidence Routing + Sweep	**0.8043**	**68.33%**	**63.33%**	**53.33%**	**46.67%**
Full Pure WSP (End-to-End)	—	—	—	—	**80.00%**

Table 3: **Decoupled Stage Ablation** isolating Stage III (Layer 35 FSM) from Stages I–II on unseen test dialogues (`Qwen3-8B`).

Ablation Variant	Verdict	Format	Consist.	Hold	Yield	Synth
1. Unmodified Base (Free Gen)	47.3%	88.9%	77.8%	59.2%	90.0%	75.5%
2. **Stage III Only (Base + FSM)**	69.5%	**100.0%**	**95.8%**	87.3%	51.7%	74.0%
3. Stages I+II Only (No Stage III)	67.5%	0.0%*(96t)*	82.5%	69.1%	89.1%	75.5%
4. Stages II+III Only (Suite `f`)	78.5%	**100.0%**	98.5%	**87.1%**	69.9%	77.1%
5. **Full Stages I+II+III (Ours)**	**86.8%**	**100.0%**	**98.9%**	86.3%	**75.1%**	**78.1%**

`(correct)` (`Yield_path`), or `(match)` (`Consensus_path`, when both *Thesis* and *Antithesis* agree and are valid).

2. **Step-2 Domain-Conditional Sub-Branching inside `(match)` at `[Domain`**: Once on the `(match):\n[Domain` path, Layer 35 executes a *second conditional branch* based on problem domain: emitting `=Math] Algebraic check confirms derivation;\n_final` for math/symbolic tasks (`num`, `value`) versus `=Logic) Conceptual review validates conclusion/\n_final` for commonsense/MCQ tasks (`bool`, `mcq`).

On 80 unseen `half=1` dialogues, Program 2 achieves a 98.75% (79/80) **exact state transition success rate**, 100.00% (34/34) **Step-2 conditional sub-branch accuracy**, 86.25% final answer accuracy, and 0.00% (0/327) **false triggers** across all 327 single-turn benchmark questions.

Table 4: **Cross-Program Generalization (Program 2: 4-Path Hierarchical 3-Way + 2-Subway State Machine)** compiled into 50 dormant Layer 35 neurons of `Qwen3-8B`, evaluated on 80 unseen `half=1` dialogues and 327 single-turn benchmark prompts.

Program 2 Metric on Unseen Test Inputs	Result
Compiled Dormant Neurons in Layer 35 / Layers [24..30]	50 / 24 per layer
Unseen Multi-Step State Transition Success Rate	**98.75%** (79/80)
Active Path Distribution (`Hold` / `Yield` / `Math` / `Logic`)	37 / 8 / 20 / 14
Step-2 Conditional Sub-Branch Accuracy (`=Math` vs. `=Logic`))	**100.00%** (34/34)
Unseen Final Answer Accuracy (80 dialogues)	**86.25%** (69/80)
False-Trigger Rate on 327 Single-Turn `thesis_prompt` Tasks	**0.00%** (0/327)

5 Conclusion

By decoupling upper-middle-layer evidence routing from terminal-layer (MLP_{L-1}) state-machine execution, Pure In-Weight WSP compiles multi-step conditional programs directly into $\sim 0.035\%$ of dormant Transformer weights in 30 seconds with zero gradient steps, outperforming reinforcement learning and constant-bias baselines across 3,780 dialogues on `Qwen3` and `Llama-3.1`.

References

[1] Benchmarks Don't Talk Back: Learning Through Self-Disagreement. *arXiv preprint arXiv:2611.17563*, 2026.

[2] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[3] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations (ICLR)*, 2023.

[4] Andy Zou et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.