hexagon:2611.00955v2

Norm-Preserving Spherical Column Rotation Mitigates Greedy Reasoning Loops Across Hybrid and Transformer Models

Igor M. Zhuk *igor.m.zhuk@gmail.com*
Independent Researcher
ORCID: 0009-0003-1670-8388

Abstract

Compact reasoning language models often stall in repetitive verification inside `<think>...</think>` scratchpads, exhausting generation budgets of up to 32,768 new tokens under deterministic greedy decoding ($T = 0$). Recent training-free weight-space programming (MIT-p50c) on the hybrid recurrent-attention model Qwen3.5-0.8B raises MATH-500 accuracy from 7.9% to 48.1% under its published sampling setup ($T = 0.6$) with an external presence penalty of 1.5. Double-precision inspection shows that over 99.999% of MIT-p50c update energy in each edited matrix lies in its first singular mode while edited column norms grow by up to 114.3%. Under greedy decoding without presence penalties ($T = 0$), its pooled pass rate across five benchmarks (1,128 prompts: four accuracy suites plus a 64-item formal completion filter) drops to 10.8% with 85.1% non-EOS cutoffs. We study **SGNE-CC** (*Spherical Gated Norm-preserving Edit, Column-Constrained*), which rotates selected output columns on their pretrained Euclidean spheres in BF16 while keeping the standard forward pass. In **Study I** on Qwen3.5-0.8B (752.4M text parameters; 18 GatedDeltaNet and six attention layers), **SGNE-CC v1** rotates 1,984 output columns (within 0.163% BF16 norm deviation) plus one decay-gate row. Pure-greedy pooled pass rate rises from 1.4% (Base) and 10.8% (MIT-p50c) to 39.8% (18.8% cutoffs), leading on all five suites, whereas under presence-penalized sampling SGNE-CC v1 trails MIT-p50c in pass rate (44.7% vs. 52.0%) at 4.1× shorter mean length (3.7k vs. 15.0k tokens). In **Study II**, to test whether recurrent layers or gate edits are needed on a pure Transformer, **SGNE-CC v2** applies a closed-form spherical rotation to 205 `mlp.down_proj` columns (0.018% of weights; within 0.030% BF16 norm deviation) on DeepSeek-R1-Distill-Qwen-1.5B with no presence penalty. Across 1,054 problems on five benchmarks, greedy accuracy rises from 60.5% to 69.5% ($+166/-71$, $p = 6.07 \times 10^{-10}$), unclosed `<think>` traces fall from 28.9% to 3.9% (305 to 41), and mean length drops from 10.8k to 4.0k tokens. At $T = 0.6$, accuracy is 72.3% vs. 73.3% (762 vs. 773 solved, $p = 0.447$) with 33.6% fewer tokens. Single-factor ablations isolating norm restoration are left to future work.

1 Introduction

Compact reasoning language models trained via distillation or reinforcement learning exhibit a recurring long-horizon failure mode: after deriving a candidate answer early inside a `<think>...</think>` block, generation repeatedly re-checks edge cases until hitting the token ceiling (Chen et al., 2024; DeepSeek-AI, 2025). At a budget of 32,768 new tokens, this failure mode is most frequent under deterministic greedy decoding ($T = 0$), where stochastic sampling cannot interrupt repetitive trajectories. Reaching the 32,768-token cap, emitting an end-of-sequence (EOS) token, and closing the `</think>` tag are distinct events and are recorded separately throughout this study.

Inference-time steering methods shorten reasoning chains by projecting or transporting hidden activations at each decoding step (Huang et al., 2025; Li et al., 2026) or by rotating activations on hyperspheres (You

et al., 2026; Dang & Ngo, 2026), which requires runtime hooks or auxiliary modules. Conversely, Takehi et al. (2026) introduced training-free *weight-space programming* and released MIT-p50c, a static edit of 128 `mlp.down_proj` columns across two layers of `Qwen/Qwen3.5-0.8B` (Qwen Team, 2026) that retains the standard forward pass. Under the MIT release's decoding configuration ($T = 0.6$, `top_p` $= 0.95$, `top_k` $= 20$, and a presence penalty of 1.5), MIT-p50c raises accuracy on the 500-problem MATH-500 subset (Hendrycks et al., 2021) from 7.9% to 48.1% while lowering 32,768-token truncations from 93.0% to 33.1%.

Evaluating a static weight edit only under a presence penalty of 1.5 (which subtracts 1.5 from the logit of any token already generated in the response, alongside temperature and top-k/top-p sampling) mixes intrinsic weight behavior with decoder-level repetition control. Moreover, Qwen3.5-0.8B is a hybrid architecture in which 18 of 24 layers (75%) are recurrent GatedDeltaNet linear-attention blocks (Yang et al., 2024) and six are grouped-query attention (GQA) layers. Examining a hybrid recurrent model and a pure Transformer in one paper addresses three questions:

1. **Greedy decoding vs. penalized sampling:** Do static weight edits suppress 32,768-token reasoning loops under pure greedy decoding ($T = 0$, presence penalty 0), and how does model ranking change under presence-penalized sampling ($T = 0.6$, presence penalty 1.5)?

2. **Weight-space geometry:** How do rank-1 additive updates that inflate edited column norms compare empirically against spherical rotations that preserve pretrained Euclidean column norms up to BF16 rounding?

3. **Closed-form pure-Transformer transfer:** Does spherical column rotation require recurrent GatedDeltaNet layers or gate edits, or does a closed-form rotation of selected feed-forward output columns suffice on a pure Transformer?

Contributions. Across 2,182 benchmark evaluations per decoding regime (1,128 in Study I and 1,054 in Study II, spanning 1,458 distinct benchmark items at a 32,768-new-token ceiling), we establish three results:

- **Weight-space audit and regime crossover on Qwen3.5-0.8B (Study I, Sections 2–3):** FP64 inspection shows that over 99.999% of MIT-p50c update energy per edited matrix lies in its first singular mode and edited column norms grow by up to 114.3%. Comparing MIT-p50c against **SGNE-CC v1** (1,984 rotated output columns within 0.163% BF16 norm deviation plus one decay-gate row) across five suites reveals a full ranking reversal between the two decoding setups. Under greedy decoding ($T = 0$, no penalty), SGNE-CC v1 leads MIT-p50c on all five suites (39.8% vs. 10.8% pooled pass rate; 18.8% vs. 85.1% cutoffs). Under MIT-p50c sampling ($T = 0.6$, presence penalty 1.5), MIT-p50c leads in pooled pass rate (52.0% vs. 44.7%), while SGNE-CC v1 generates 4.1× fewer tokens (3.7k vs. 15.0k).

- **Closed-form feed-forward edit on a pure Transformer (Study II, Sections 2 and 4):** Guided by an attention-head and SwiGLU gate audit on `DeepSeek-R1-Distill-Qwen-1.5B`, **SGNE-CC v2** applies a closed-form spherical rotation (Eq. 1) to 205 `mlp.down_proj` columns (0.018% of parameters; within 0.030% BF16 norm deviation) toward the centered `</think>` unembedding direction, with zero attention or gate edits and no presence penalty. Across 1,054 problems on five benchmarks, greedy accuracy rises by 9.0 percentage points from 60.5% to 69.5% ($+166/-71$, $p = 6.07 \times 10^{-10}$), unclosed `<think>` responses drop from 28.9% to 3.9%, and mean length falls from 10.8k to 4.0k tokens.

- **Multi-stage calibration exclusions and failure boundaries (Sections 4.1 and 6):** Greedy gains hold on MATH-500 after excluding late calibration, attention-audit, and early newline-audit items ($N = 474$, 421, 406, and 392) and across 554 untouched code, science, and logic tasks (56.1% to 63.4% at $T = 0$, $p = 1.08 \times 10^{-3}$). We also report the 71 lost greedy solutions, duplicate `</think>` tags, T4 FP16 vs. BF16 parity, and unisolated causal factors.

Table 1: **Weight-space geometry on `Qwen/Qwen3.5-0.8B`** (FP64 audit of edited BF16 matrices; 752.4M text parameters). MIT-p50c concentrates over 99.999% of update energy per matrix in a single singular mode and inflates column norms by up to 114.26%, whereas SGNE-CC v1 keeps column norms within 0.163%.

Weight-Space Property	Base	MIT-p50c	SGNE-CC v1 (Ours)
Scalars in edited region (bitwise changed in BF16)	0	131,072 (130,188)	2,032,640 (1,993,382)
Share of 752,393,024 text parameters	0%	0.017%	0.27%
Edited `mlp.down_proj` output columns	0	128 (layers 15, 19)	802 (15 MLP layers)
Edited GatedDeltaNet `out_proj` columns	0	0 (0/18 layers)	1,182 (9/18 layers)
Edited attention `o_proj` columns (6 GQA layers)	0	0	0
Edited GatedDeltaNet decay-gate rows (of 288)	0	0	1 (layer 16, row 12)
Frobenius norm $\|\Delta W\|_F$ / mean energy per column	0 / 0	2.24 / 0.0392	3.57 / 0.0064
Column rotation angle (median / max)	0°	27.1° / 87.3°	8.8° / 93.7°
Top singular mode energy share (L15 / L19 / L23)	n/a	99.9995%/99.9991%/n/a	67.6%/73.4%/55.9%
Edited-to-base column norm ratio $\|w'_j\|_2/\|w_j\|_2$	1.0000	[1.0028, 2.1426]	[0.9984, 1.0016]
Maximum column norm deviation after BF16	0.00%	+**114.26**% (layer 19)	**0.16**% (0.1625%)

2 Weight-Space Geometry and Spherical Column Rotation

Norm distortion under additive rank-1 updates. Let $W \in \mathbb{R}^{d\times m}$ denote a residual-stream output matrix such as `mlp.down_proj`, where column $w_j = W[:, j] \in \mathbb{R}^d$ is multiplied by upstream activation $a_j(x)$ to contribute $a_j(x)w_j$. Under an additive rank-1 edit $W' = W + uv^\top$, column j changes in squared Euclidean norm by

$$\|w_j + v_j u\|_2^2 - \|w_j\|_2^2 = 2v_j w_j^\top u + v_j^2\|u\|_2^2,$$

coupling directional rotation with column-norm expansion or contraction prior to downstream RMSNorm layers. Rotating selected columns on their pretrained Euclidean spheres instead preserves $\|w'_j\|_2 = \|w_j\|_2$ up to BF16 rounding while reorienting their residual-stream directions. Preserving individual column norms does not by itself preserve inter-column angles, layer output norms, or output logits.

2.1 Geometry of MIT-p50c vs. SGNE-CC v1 on Qwen3.5-0.8B

Both MIT-p50c and SGNE-CC v1 keep the architecture, tokenizer, and forward pass of `Qwen/Qwen3.5-0.8B` unchanged (752,393,024 text parameters across layers 0–23: 752,390,432 stored in BF16 and 2,592 normalization scalars in FP32). Table 1 reports the FP64 audit of both checkpoints against Base (`apply_dna_patch.py --compare`).

MIT-p50c is rank-1 per matrix with up to 114% column-norm growth, whereas SGNE-CC v1 preserves norms within 0.163%. In layers 15 and 19 of MIT-p50c, 99.9995% and 99.9991% of ΔW energy lies in the first singular mode, and edited `mlp.down_proj` column norms increase by up to +70.59% in layer 15 and +114.26% in layer 19. SGNE-CC v1 rotates 1,984 output columns (802 feed-forward and 1,182 GatedDeltaNet `out_proj` columns across 15 MLP and nine recurrent layers, leaving all six GQA layers untouched) on their Euclidean spheres with median angle 8.8° and maximum BF16 norm deviation 0.1625%, plus one decay-gate row in layer 16 (row 12, cosine 0.9866, norm 1.0286 to 1.0452). The archive includes the final 3.13 MiB BF16 patch (`sgne_cc_dna_delta.npz`); reproducing the offline search that selected its per-column angles would require external search logs. Because MIT-p50c and SGNE-CC v1 differ simultaneously in column-norm preservation, update rank, edited module types, and the layer-16 gate row, Study II tests an explicit closed-form spherical edit restricted to `mlp.down_proj` on a pure Transformer.

2.2 Closed-Form Spherical Rotation (SGNE-CC v2) on DeepSeek-R1-Distill-Qwen-1.5B

To test whether recurrent blocks or gate edits are required on a non-hybrid model, we evaluate `deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B` (28 GQA layers, 1,777,088,000 BF16 parameters, hidden dimension $d = 1536$, SwiGLU intermediate size 8,960).

Attention-head localization vs. feed-forward-only editing. Auditing all 336 query heads (28×12) across 32 looped and 32 cleanly terminated MATH-500 traces shows that all 12 heads with the largest increase in self-repetition attention (+29.3 to +66.8 percentage points) belong to the first key-value group (KV Group 0, heads 0–5). In layer 23, head 4, prompt attention drops from 31.4% to 0.3%, and self-repetition attention reaches 67%–83% across the top four heads ranked by repetition-attention gain. In development probes on non-looping positions, however, next-token agreement with Base was 98.8% for a feed-forward-only edit versus 77.5% for an intermediate composite edit that also modified `gate_proj` and attention output projections. SGNE-CC v2 therefore leaves all attention, normalization, `gate_proj`, and `up_proj` tensors unchanged.

Gate-conditioned column selection and closed-form rotation. In SwiGLU layer ℓ, column $w_j = W_{\text{down}}^{(\ell)}[:, j] \in \mathbb{R}^{1536}$ contributes $\text{SiLU}(g_j)u_j w_j$, gated by g_j and u_j (for example, $\text{SiLU}(-1) \approx -0.269$). Rather than editing `gate_proj`, we select columns whose pretrained gates separate structural boundary tokens during repetition from ordinary reasoning tokens (using 23 early newline-calibration items and 26 late punctuation/sizing items from MATH-500, all audited under explicit exclusion splits in Section 4.1; zero of the 554 non-math tasks were inspected):

1. **Newline-boundary columns (135 columns, layers 18–27):** active at `\n` and `\n\n` boundaries during repetition, attenuated inside clauses and after `</think>`.
2. **Punctuation columns (70 columns, layers 19–27):** active on '`.`', '`?`', and '`!`' inside line-break-free repeating paragraphs (mean gate input +0.85 during looping vs. −1.41 on non-looping punctuation and −1.81 after `</think>`).

Let $u_{\text{ET}} = \tilde{u}/\|\tilde{u}\|_2 \in \mathbb{R}^{1536}$ denote the unit reasoning-end direction computed from the final RMSNorm-weighted centered unembedding $\tilde{u} = \gamma_{\text{norm}} \odot \big(W_{\text{lm}}[\texttt{</think>}] - \bar{W}_{\text{lm}}\big)$, and let $s_j = \text{sign}\big(\overline{\text{SiLU}(g_j)u_j}\big) \in \{\pm 1\}$ be column j's empirical activation sign during looping (111 positive and 94 negative across the 205 columns). Each selected column is rotated on its pretrained sphere of radius $\|w_j\|_2$ with gain $\beta = 2$ and rounded to BF16:

$$w_j' = \text{BF16}\left(\|w_j\|_2 \cdot \frac{w_j + \beta s_j \|w_j\|_2 u_{\text{ET}}}{\big\|w_j + \beta s_j \|w_j\|_2 u_{\text{ET}}\big\|_2} \right), \qquad \beta = 2. \tag{1}$$

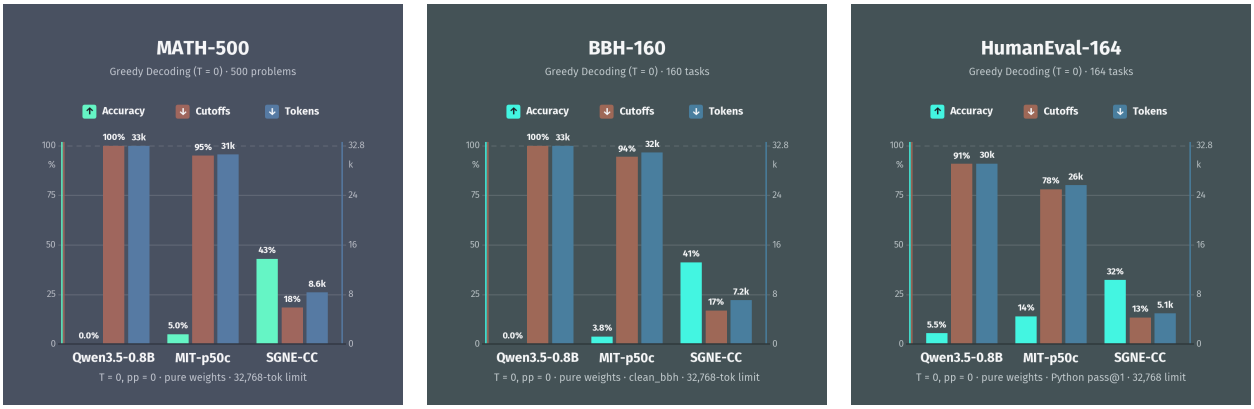
Across all 205 edited columns (314,880 scalars, or 0.0177% of model weights; 314,418 bitwise changed in BF16), rotation angles span 59.9°–66.6°, the FP64 mean edited-to-base norm ratio is 0.999993, and individual column norm ratios lie in [0.99978, 1.00030] (maximum deviation 0.030%). The released patch (`FINAL_SGNE_CC_V2_205_PATCH.npz`) was assembled with MLX/FP32 reductions (314,872/314,880 BF16 scalars identical to pure FP32 evaluation of Eq. 1 and 314,858/314,880 to pure FP64, maximum relative vector difference 3.13×10^{-4}); applying the patch is bitwise deterministic, whereas recomputing column indices and signs from scratch requires the calibration activation arrays.

3 Study I: Hybrid Recurrent-Attention Model (Qwen3.5-0.8B)

Experimental setup. We evaluate Base Qwen3.5-0.8B, MIT-p50c, and SGNE-CC v1 with reasoning enabled, one pass per prompt (seed = 0), and a 32,768-new-token generation budget (distinct from the 262,144-token positional window) across five suites totaling 1,128 prompts per regime: MATH-500 (500; `HuggingFaceH4/MATH-500`) (Hendrycks et al., 2021), multilingual MGSM-240 (40 story problems across six languages: RU, ZH, JA, FR, ES, DE) (Shi et al., 2023), BBH-160 (eight logic subtasks ×20, normalized `clean_bbh` grading) (Suzgun et al., 2023), HumanEval-164 (Python unit tests) (Chen et al., 2021), and MultiSpeech-64 (eight genres × eight languages, automated formal completion and repetition filter). Runs use batched PyTorch inference (including shards executed with FP16 Tensor Core matmul and FP32 normalizations on T4; Section 6) under two configurations: **Greedy** ($T = 0$, `top_p` = 1.0, `top_k` = 0, presence penalty 0) and **Card** ($T = 0.6$, `top_p` = 0.95, `top_k` = 20, presence penalty 1.5 on generated tokens, matching the MIT-p50c release configuration). Math solutions are graded symbolically via `math-verify`, requiring closed `</think>`, normal EOS termination, and an exact `\boxed{}` expression.

Table 2: **Study I benchmark results on `Qwen/Qwen3.5-0.8B`** across five suites and two decoding configurations (32,768-token budget). **Pass** (↑, %) denotes accuracy on rows 1–4 and formal completion filter pass rate on MultiSpeech-64; **Cut** (↓, %) denotes non-EOS 32,768-token truncations; **Tok** (↓) denotes mean generated tokens. Best per row in **bold**, second-best underlined. Paired two-sided p-values compare SGNE-CC v1 against MIT-p50c (exact McNemar; 40-block clustered sign-flip on MGSM-240).

Suite (N)	Regime	Base Qwen3.5-0.8B			MIT-p50c			SGNE-CC v1 (Ours)			v1 vs. MIT (+/−, p)
		Pass↑	Cut↓	Tok↓	Pass↑	Cut↓	Tok↓	Pass↑	Cut↓	Tok↓	
MATH-500 (500)	Greedy	0.0	100.0	32.8k	5.0	95.0	31.4k	**43.0**	**18.4**	**8.6**k	$+195/-5$, 3.24×10^{-51}
	Card	7.8	92.2	30.6k	**51.8**	28.2	21.1k	47.0	**1.6**	**5.3**k	$+56/-80$, 0.048
MGSM-240 (240)	Greedy	2.1	97.5	32.0k	21.3	70.8	24.0k	**38.3**	**20.4**	**8.2**k	$+54/-13$, 1.67×10^{-6}
	Card	25.8	66.7	23.6k	**57.1**	3.8	8.6k	41.7	**2.1**	**3.1**k	$+15/-52$, 9.04×10^{-5}
BBH-160 (160)	Greedy	0.0	100.0	32.8k	3.8	94.4	31.7k	**41.3**	**16.9**	**7.2**k	$+61/-1$, 2.73×10^{-17}
	Card	1.9	96.3	31.9k	**50.6**	14.4	16.9k	41.3	**0.6**	**2.4**k	$+20/-35$, 0.058
HumanEval (164)	Greedy	5.5	90.9	29.9k	14.0	78.0	26.3k	**32.3**	**13.4**	**5.1**k	$+37/-7$, 5.30×10^{-6}
	Card	15.9	35.4	12.6k	**31.1**	13.4	8.8k	28.0	**0.0**	**1.1**k	$+13/-18$, 0.473
MultiSpeech (64)	Greedy	3.1	96.9	31.8k	26.6	56.3	20.2k	**35.9**	**34.4**	**13.0**k	$+11/-5$, 0.210
	Card	67.2	29.7	11.5k	**90.6**	**0.0**	**2.5**k	89.1	3.1	2.9k	$+4/-5$, 1.000
Pooled (1,128)	**Greedy**	1.4	98.0	32.1k	10.8	85.1	28.5k	**39.8**	**18.8**	**8.0**k	449 vs. 122 passed
Pooled (1,128)	**Card**	15.3	75.5	25.6k	**52.0**	17.3	15.0k	44.7	**1.4**	**3.7**k	504 vs. 586 passed



(a) MATH-500 ($N = 500$, **Greedy**) **(b) BBH-160** ($N = 160$, **Greedy**) **(c) HumanEval-164** ($N = 164$, **Greedy**)

Figure 1: **Greedy decoding without presence penalties separates SGNE-CC v1 from MIT-p50c on Qwen3.5-0.8B.** Across (a) MATH-500, (b) BBH-160, and (c) HumanEval-164 at a 32,768-token ceiling ($T = 0$, presence penalty 0), Base and MIT-p50c reach the token limit without EOS on 78%–100% of prompts, whereas SGNE-CC v1 reaches 32.3%–43.0% accuracy at 5.1k–8.6k mean tokens. On the cards, `SGNE-CC` denotes SGNE-CC v1 and `pp = 0` denotes presence penalty 0; mint (`Accuracy` ↑, left axis, %), brown (`Cutoffs` ↓, non-EOS truncations, %), and blue (`Tokens` ↓, right axis, k) bars round the exact values in Table 2.

Switching from penalized sampling to greedy decoding reverses model ranking across all five suites. Table 2 and Figure 1 compare the two decoding configurations (between which temperature, top-k/top-p truncation, and presence penalty change jointly; isolating the standalone effect of presence penalty at fixed temperature would require an additional run). Under pure greedy decoding ($T = 0$, no penalty), MIT-p50c solves 5.0% (25/500) on MATH-500 with 95.0% cutoffs, frequently deriving the right intermediate value early inside `<think>` and then repeating an edge-case check until token 32,768, whereas SGNE-CC v1 solves 43.0% (215/500, 18.4% cutoffs, 8.6k tokens; $+195/-5$, $p = 3.24\times10^{-51}$). Under lenient fallback extraction from truncated 32k drafts, SGNE-CC v1 leads 44.2% vs. 18.6% (221 vs. 93; Appendix A). On the 300 MATH-500 items disjoint from the ≤ 200 items inspected during angle selection (Holdout-300), greedy

Table 3: **Study II benchmark results on `DeepSeek-R1-Distill-Qwen-1.5B`** (BF16, presence penalty 0, 32,768-token budget). **Acc.** (↑, %) is Pass@1 accuracy; **Uncl.** (↓, %) is the share of responses lacking a closing `</think>` tag; **Tok** (↓) is mean generated tokens; **C/W/U** gives exact counts of Correct, Wrong after `</think>`, and Unclosed responses. Best accuracy, unclosed rate, and token length per row in **bold**.

Benchmark (N)	Temp.	Base DeepSeek-R1-1.5B				SGNE-CC v2 (Ours, 205 cols)				Δ Acc.	McNemar ($+/-$, p)
		Acc.↑	Uncl.↓	Tok↓	C / W / U	Acc.↑	Uncl.↓	Tok↓	C / W / U		
MATH-500 (500)	$T=0$	65.4	27.8	11.0k	327/34/139	**76.4**	**3.2**	**4.7**k	382/102/16	+11.0	$+74/-19$, 7.72×10^{-9}
	$T=0.6$	80.0	3.0	5.8k	400/85/15	**82.0**	**0.4**	**4.3**k	410/88/2	+2.0	$+33/-23$, 0.229
MBPP-120 (120)	$T=0$	53.3	37.5	13.1k	64/11/45	**67.5**	**5.8**	**3.6**k	81/32/7	+14.2	$+21/-4$, 9.11×10^{-4}
	$T=0.6$	**67.5**	3.3	4.8k	81/35/4	64.2	**0.0**	**2.2**k	77/43/0	−3.3	$+7/-11$, 0.481
HumanEval (164)	$T=0$	52.4	38.4	13.8k	86/15/63	**53.7**	**10.4**	**6.2**k	88/59/17	+1.2	$+20/-18$, 0.871
	$T=0.6$	65.9	4.3	5.8k	108/49/7	**68.9**	**1.2**	**3.3**k	113/49/2	+3.0	$+16/-11$, 0.442
ARC-Chall. (150)	$T=0$	52.7	22.7	7.7k	79/37/34	**59.3**	**0.0**	**0.8**k	89/61/0	+6.7	$+28/-18$, 0.184
	$T=0.6$	**57.3**	0.7	1.5k	86/63/1	56.7	**0.0**	**1.1**k	85/65/0	−0.7	$+18/-19$, 1.000
BBH Logic (120)	$T=0$	68.3	20.0	7.4k	82/14/24	**77.5**	**0.8**	**2.4**k	93/26/1	+9.2	$+23/-12$, 0.090
	$T=0.6$	72.5	1.7	2.5k	87/31/2	**73.3**	**0.0**	**1.5**k	88/32/0	+0.8	$+18/-17$, 1.000
Pooled (1,054)	$T=0$	60.5	28.9	10.8k	638/111/305	**69.5**	**3.9**	**4.0**k	733/280/41	**+9.0**	$+166/-71$, $\mathbf{6.07\times10^{-10}}$
Pooled (1,054)	$T=0.6$	72.3	2.8	4.7k	762/263/29	**73.3**	**0.4**	**3.1**k	773/277/4	+1.0	$+92/-81$, 0.447

accuracy is 0.0% for Base, 6.7% for MIT-p50c, and 51.3% for SGNE-CC v1 ($+138/-4$, $p = 6.00\times10^{-36}$); on the SHA-256-sampled FreshBlind-49 subset, solved counts are 0/49, 2/49, and 24/49 ($p = 4.77\times10^{-7}$). Out-of-domain, greedy SGNE-CC v1 reaches 41.3% (66/160) on BBH-160 vs. 3.8% for MIT-p50c (strict boxed counts: 0/6/51 at $T=0$ and 1/66/54 at $T=0.6$) and achieves the study's highest HumanEval-164 score of 32.3% (53/164, 12.5× shorter `<think>`), surpassing MIT-p50c under Card sampling (31.1%, 51/164).

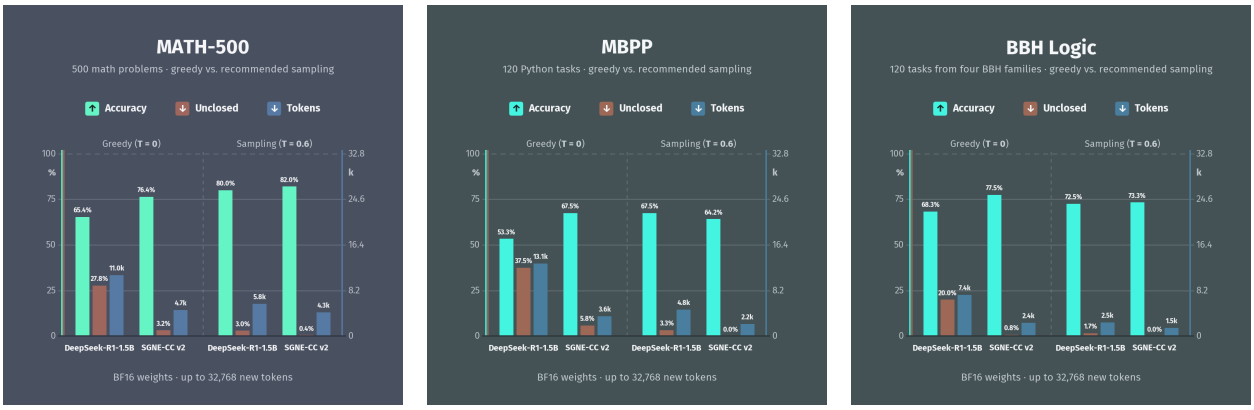
Under presence-penalized sampling, MIT-p50c leads in pass rate while SGNE-CC v1 is 4.1× shorter. Under the Card configuration ($T=0.6$, presence penalty 1.5), MIT-p50c outperforms SGNE-CC v1 on multilingual MGSM-240 (57.1% vs. 41.7%, 40-block clustered $p = 9.04\times10^{-5}$) and in direct PyTorch runs on MATH-500 (51.8% vs. 47.0%, nominal $p = 0.048$; compared against MIT's external vLLM reference of 48.2% run-1 and 48.1% four-run mean, $+65/-71$, $p = 0.668$). Across all 1,128 prompts in this regime, SGNE-CC v1 generates 4.1× fewer tokens (3.7k vs. 15.0k; exact means 3,681 vs. 14,985) and reduces non-EOS cutoffs by 12× from 17.3% (195/1,128) to 1.4% (16/1,128).

4 Study II: Closed-Form Transfer to a Pure Transformer

Experimental setup. We evaluate Base `DeepSeek-R1-Distill-Qwen-1.5B` and SGNE-CC v2 (205 `mlp.down_proj` columns, Eq. 1) in BF16 on vLLM 0.34.0 with presence penalty fixed to 0 across all runs and a budget of 32,768 new tokens across five benchmarks (1,054 problems): MATH-500 (500), MBPP-120 (120; prompts show the first three assertions per task, or 360 of the 365 assertions executed in scoring) (Austin et al., 2021), HumanEval-164 (164, hidden unit tests), ARC-Challenge-150 (150) (Clark et al., 2018), and BBH Logic-120 (120 across four logic families). Both models are compared at $T=0$ and $T=0.6$ (`top_p` $= 0.95$, seed $= 42$, one pass per prompt), plus $T=0.3$ operating-point runs for SGNE-CC v2 (Base was not evaluated at $T=0.3$). Any trace without a closing `</think>` tag is scored incorrect.

4.1 Calibration-Exclusion Splits and Out-of-Domain Transfer

Greedy gains persist after excluding late, attention-audit, and early calibration items ($N = 474$, 421, 406, and 392). On full MATH-500 at $T=0$ (Table 3, Figure 2), rotating 205 columns lowers unclosed `<think>` traces by 8.7× from 27.8% to 3.2% (139 to 16; non-EOS cap hits likewise drop from 139 to 16), cuts mean output from 11.0k to 4.7k tokens, and raises greedy accuracy by 11.0 percentage points from 65.4% to 76.4% ($+74/-19$, $p = 7.72\times10^{-9}$). Under strict boxed-only scoring without final-line fallback, solved counts still rise from 327 to 376 at $T=0$ and from 400 to 409 at $T=0.6$. Excluding the 26 late punctuation/sizing problems ($N = 474$) gives 65.4% vs. 76.6% (310 to 363 solved; $+70/-17$, $p = 8.35\times10^{-9}$), and additionally



(a) MATH-500 ($N = 500$, $T = 0$ **vs.** 0.6) **(b) MBPP-120** ($N = 120$, $T = 0$ **vs.** 0.6) **(c) BBH Logic** ($N = 120$, $T = 0$ **vs.** 0.6)

Figure 2: **Closed-form rotation of 205 feed-forward columns raises greedy accuracy across MATH-500, MBPP-120, and BBH Logic-120 on DeepSeek-R1-Distill-Qwen-1.5B.** At $T = 0$ (presence penalty 0, 32,768-token budget), SGNE-CC v2 lowers unclosed `<think>` rates (`Unclosed` ↓, brown bars, left axis, %) from 20.0%–37.5% to 0.8%–5.8% and raises greedy accuracy (`Accuracy` ↑, mint bars, left axis, %) by +11.0, +14.2, and +9.2 percentage points on (a)–(c). At $T = 0.6$, mean token counts (`Tokens` ↓, blue bars, right axis, k) fall by 26%–54%, while accuracy changes are +2.0 pp on MATH-500 (400 to 410), −3.3 pp on MBPP-120 (81 to 77, $p = 0.481$), and +0.8 pp on BBH Logic-120 (87 to 88; Table 3).

excluding the 53 distinct problems from the 64-trace head audit ($N = 421$) gives 65.6% vs. 76.2% (276 to 321 solved; $+60/-15$, $p = 1.59 \times 10^{-7}$). Because the 135 newline columns also used mean gate activations from 23 early-stage problems (15 of which remained inside $N = 421$), excluding those 15 items as well ($N = 406$) yields 66.0% vs. 75.9% at $T = 0$ (268 to 308 solved; $+54/-14$, $p = 1.11 \times 10^{-6}$; at $T = 0.6$: 327 vs. 333, $p = 0.461$). Excluding the entire 40-item early candidate pool ($N = 392$) yields 66.3% vs. 75.5% (260 to 296 solved; $+49/-13$, $p = 4.82 \times 10^{-6}$; explicit item IDs in the supplementary verifier). Comparing SGNE-CC v2 at $T = 0.3$ (78.0%, 390/500, four unclosed) with Base at $T = 0.6$ (80.0%, 400/500, 15 unclosed) contrasts two operating points differing in both weights and temperature ($+29/-39$, $p = 0.275$).

Transfer across 554 untouched code, science, and logic tasks. None of the four non-math benchmarks were used during head inspection or column selection. Pooled across all 554 non-math tasks at $T = 0$, accuracy increases from 56.1% to 63.4% (311 to 351 solved; $p = 1.08 \times 10^{-3}$; at $T = 0.6$: 362 vs. 363 solved). Across all 1,054 greedy tasks, unclosed `<think>` responses, non-EOS cap hits, and high 6-gram repetition (≥ 0.50) fall from 305, 300, and 302 on Base to 41, 38, and 75 on SGNE-CC v2 (pooled mean tokens fall from 10,782 to 3,983 at $T = 0$ and from 4,699 to 3,119 at $T = 0.6$). On MBPP-120 at $T = 0$, accuracy rises by 14.2 percentage points from 53.3% to 67.5% (64 to 81/120; $+21/-4$, $p = 9.11 \times 10^{-4}$), matching Base at $T = 0.6$ while generating 3.7× fewer tokens than greedy Base (13.1k vs. 3.6k). On HumanEval-164 at $T = 0$, unclosed traces drop from 38.4% to 10.4% (63 to 17) and 49 of 53 newly closed drafts parse as valid Python via `ast.parse`, while unit-test pass counts change by +2 net ($+20/-18$, $p = 0.871$; syntax validity does not imply functional correctness). On ARC-Challenge-150 (where the grader checks `\boxed{}` first and falls back to answer-letter patterns; 87/89 correct SGNE-CC v2 responses at $T = 0$ and 85/85 at $T = 0.6$ used fallback extraction), unclosed traces drop to 0.0% across all temperatures, greedy length falls 9.8× (7.7k to 0.8k), and SGNE-CC v2 at $T = 0.3$ scores 66.0% (99/150) vs. 57.3% (86/150) for Base at $T = 0.6$ ($+31/-18$, $p = 0.085$). On BBH Logic-120 at $T = 0$, unclosed traces drop from 20.0% to 0.8% (24 to 1), length falls 3.1×, and accuracy moves from 68.3% to 77.5% ($+23/-12$, $p = 0.090$).

5 Related Work

Overthinking and inference-time reasoning control. Reasoning models frequently continue generating redundant verification rounds long after their internal answer distribution has stabilized (Chen et al., 2024;

Table 4: **Cross-architecture summary of SGNE-CC v1 and SGNE-CC v2** at a 32,768-new-token budget.

Property (BF16 weights, 32,768-token budget)	Study I (v1): Qwen3.5-0.8B	Study II (v2): DeepSeek-R1-1.5B
Architecture	Hybrid (18 GatedDeltaNet + 6 GQA), 752M	Pure Transformer (28 GQA layers), 1.78B
Edited parameters	1,984 cols + 1 gate row (0.27%, 3.13 MiB)	205 `mlp.down_proj` cols (**0.018**%, 0.58 MiB)
Max BF16 column norm deviation	0.163% ([0.9984, 1.0016])	**0.030**% ([0.99978, 1.00030])
MATH-500 greedy ($T = 0$, no penalty)	0.0% → **43.0**% (cutoffs 100.0% → 18.4%)	65.4% → **76.4**% (unclosed 27.8% → 3.2%)
Pooled greedy score ($T = 0$, no penalty)	1.4% → **39.8**% pass ($N = 1{,}128$)	60.5% → **69.5**% acc. ($N = 1{,}054$, $p = 6.07 \times 10^{-10}$)
Sampling ($T = 0.6$) comparison	44.7% vs. 52.0% MIT (4.1× shorter, pen. 1.5)	72.3% → 73.3% (−33.6% tokens, no penalty)

DeepSeek-AI, 2025). Recent inference-time methods regulate reasoning length through token-budget forcing on `Wait` and `</think>` delimiters (Muennighoff et al., 2025), activation manifold projection (Huang et al., 2025), or learned flow-matching transport between verbose and concise hidden representations (Li et al., 2026). Unlike runtime interventions that alter activations at each decoding step, SGNE-CC modifies pretrained BF16 weights offline and retains the unmodified forward pass.

Norm-preserving steering vs. static weight-space editing. Additive activation steering (Turner et al., 2024) and rank-1 directional ablation (Arditi et al., 2024) alter hidden-state norms, motivating runtime norm-preserving activation rotations such as Spherical Steering (You et al., 2026) and Selective Steering (Dang & Ngo, 2026). On the static weight-editing side, factual editing (Meng et al., 2022; Fang et al., 2025) and task arithmetic (Ilharco et al., 2023) apply additive weight updates. Closest to our setting, Sun et al. (2026) (Steer2Edit) convert steering vectors into rank-1 additive updates across attention W_o and MLP W_{down} components via an Elastic-Net grid search, Jayabahu & Adeleke (2026) distill a layer-18 halt vector into `DeepSeek-R1-Distill-Qwen-7B` weights by reconstructing steered activations with off-axis dimensions pinned, and Takehi et al. (2026) optimize 128 `mlp.down_proj` columns in `Qwen3.5-0.8B`. SGNE-CC differs in geometry and evaluation regime: on `DeepSeek-R1-Distill-Qwen-1.5B`, it rotates 205 gate-selected `mlp.down_proj` columns on their Euclidean spheres in closed form without adapter training or hyperparameter sweeps over attention heads, and evaluates completion and accuracy at a 32,768-token ceiling without presence penalties.

6 Synthesis, Limitations, and Conclusion

Limitations and scope boundaries. Table 4 compares both studies. On DeepSeek-R1-Distill-Qwen-1.5B, Study II shows that editing 205 feed-forward columns (15× smaller share of total parameters than Study I, 0.018% vs. 0.27%, and 9.7× fewer edited columns) suffices without recurrent blocks or gate edits. Three boundaries govern interpretation:

1. **Unisolated single-factor causality:** Column-norm preservation and update rank are distinct geometric properties: prior to BF16 rounding, Eq. 1 writes $\Delta W = W\,\mathrm{diag}(\alpha_j - 1) + u_{\mathrm{ET}} c^\top$ (yielding multiple nonzero singular values), whereas a Householder reflection $I - 2uu^\top$ preserves every column norm with $\mathrm{rank}(\Delta W) \le 1$. Because MIT-p50c and SGNE-CC v1 differ simultaneously in column norms, update rank, edited modules, and one decay-gate row, isolating the standalone causal effect of norm restoration requires single-factor ablations (rescaling MIT-p50c's 128 columns to pretrained norms, or omitting the $\|w_j\|_2$ normalization factor in Eq. 1 on the same 205 columns, signs, and $\beta = 2$).

2. **Regime trade-offs, lost solutions, and duplicate tags:** Under presence-penalized sampling in Study I, shorter SGNE-CC v1 traces score lower than MIT-p50c on MGSM-240 (41.7% vs. 57.1%) and in direct PyTorch runs on MATH-500 (47.0% vs. 51.8%). In Study II at $T = 0$, the 166 gained solutions are accompanied by 71 lost solutions (18 closed earlier than Base, 38 closed later, 15 unclosed), so first-closure timing alone does not explain individual errors. SGNE-CC v2 also emits a second `</think>` token inside the visible answer on 84 responses at $T = 0$ and 83 at $T = 0.6$ (vs. zero and three in Base). Greedy scoring across all 1,054 tasks is insensitive to duplicate-tag parsing: slicing strictly after the last `</think>` gives 732/1,054 (69.4%, unchanged at 382/500 on

MATH-500), slicing strictly between the first and second tag gives 725/1,054 (68.8%), and scoring all 84 duplicate-tag responses as incorrect still yields 701/1,054 (66.5% vs. 638/1,054 for Base). At $T = 0.6$ in Study II, accuracy changes by -3.3 pp on MBPP-120 (81 to 77, $p = 0.481$) and pooled accuracy does not differ significantly ($p = 0.447$); the observed effect under sampling is a 33.6% reduction in generated tokens and fewer unclosed traces (29 to 4).

3. **Hardware precision, single-seed protocol, and surface filters:** Each regime evaluates one generation per prompt, nominal p-values are unadjusted for multiplicity, and $p > 0.05$ does not establish equivalence. In Study I, 2,892 of 6,768 primary trace rows record accelerator metadata (native BF16 on A100 and RTX PRO 6000 Blackwell vs. FP16 Tensor Core matmul with FP32 normalizations on T4 and BF16 on L4); a 60-trace $T = 0$ parity check gives 20/20 item agreement on Base, 19/20 on MIT-p50c, and 13/20 on SGNE-CC v1 (10/20 vs. 9/20 solved). In Study II, all 5,270 traces ran in BF16 on vLLM 0.34.0; 316 greedy Base traces were resumed from 8,192 with 24,576 continuation tokens via re-tokenization (reaching 32,769 tokens on one BBH Logic item, logged in the certificate). Finally, MultiSpeech-64 tests formal completion rather than factual fidelity: under a presence penalty of 1.5, isolated Chinese characters appear in Arabic/Russian outputs (three in MIT-p50c, two in SGNE-CC v1), four of SGNE-CC v1's five repetition failures occur in German, and on `ru/historical_portrait` SGNE-CC v1 writes "F. 1895" in place of Fridtjof Nansen while passing the formal filter.

Independent external replications. Subsequent to the initial public release of Study I (Zhuk, 2026a), two external groups independently evaluated the BF16 checkpoint `SGNE-CC v1` on separate inference stacks. Uhlemann (2026) tested `SGNE-CC v1` in `llama.cpp` (Sections 3.13–3.14), confirming its token-efficiency advantage over MIT-p50c and replicating a GatedDeltaNet-only ablation arm. On the 300-problem held-out manifest in Hugging Face `transformers`, Human-AI-Evolution (2026) replicated greedy recovery (143/300 = 47.7% strict) and reported a five-arm component ablation showing that removing the single layer-16 decay-gate row does not reduce accuracy (147/300), while zeroing the 1,182 GatedDeltaNet `out_proj` columns preserves clean termination (0% cap hits, 37.3% strict) at higher mid-trace repetition (49.0% vs. 27.7%).

Conclusion. Testing static weight edits at a 32,768-token budget under both pure greedy decoding and presence-penalized sampling separates intrinsic weight behavior from decoder heuristics. On hybrid Qwen3.5-0.8B, SGNE-CC v1 raises greedy pass rates across all five suites while trailing MIT-p50c in accuracy at 4.1× shorter output length under penalized sampling; on pure-Transformer DeepSeek-R1-Distill-Qwen-1.5B, closed-form spherical rotation of 205 feed-forward columns raises greedy accuracy by 9.0 percentage points and shortens sampled generations by 33.6% while retaining the standard forward pass.

Broader Impact Statement

Across the two models, static spherical column edits reduced mean greedy token counts by 2.7×–4.0× without gradient retraining (wall-clock latency and hardware energy consumption were not measured in these runs). Because cleaner scratchpad closure can make factually incorrect answers appear well-formed, as shown in the MultiSpeech-64 audit, downstream applications must verify factual and code outputs independently and execute model-generated code only inside isolated sandboxes.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do NOT think that much for 2+3=? On the overthinking of o1-like LLMs. *arXiv preprint arXiv:2412.21187*, 2024.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Quy-Anh Dang and Chris Ngo. Selective steering: Norm-preserving control through discriminative layer selection. *arXiv preprint arXiv:2601.19375*, 2026.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Shi Jie, Xiang Wang, Xiangnan He, and Tat-seng Chua. AlphaEdit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks Track*, 2021. (500-problem evaluation subset: `HuggingFaceH4/MATH-500`.)

Yao Huang, Huanran Chen, Shouwei Ruan, Yichi Zhang, Xingxing Wei, and Yinpeng Dong. Mitigating overthinking in large reasoning models via manifold steering. *arXiv preprint arXiv:2505.22411*, 2025.

Human-AI-Evolution. Weight-space notes and shared-rows evaluation benchmark (five-arm `SGNE-CC` greedy and ablation rows), 2026. https://github.com/Human-AI-Evolution/weight-space-notes.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

Dylan Jayabahu and Tinuade Adeleke. The halt vector: Internalizing a causal steering intervention for efficient reasoning. *arXiv preprint arXiv:2608.28859*, 2026.

Yawei Li, Benjamin Bergner, Yinghan Zhao, Vihang Prakash Patil, Bei Chen, and Cheng Wang. Steering large reasoning models towards concise reasoning via flow matching. *Transactions on Machine Learning Research (TMLR)*, 2026.

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.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

Qwen Team. *Qwen3.5-0.8B*. Model card, 2026. https://huggingface.co/Qwen/Qwen3.5-0.8B.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *International Conference on Learning Representations (ICLR)*, 2023.

Chung-En Sun, Ge Yan, Zimo Wang, and Tsui-Wei Weng. Steer2Edit: From activation steering to component-level editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL*, 2023.

Rikiya Takehi, Jayden Teoh, Alex Zhang, and Omar Khattab. Sparks of weight-space programming, 2026. MIT CSAIL research release and reference logs. https://rikiyat.github.io/blog/weight-space-programming/.

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 preprint arXiv:2308.10248v4*, 2024.

Dave Uhlemann. Knowing is not committing: Weight-space analysis and replication report (version 0.5, Sections 3.13–3.14), 2026. Zenodo preprint, https://doi.org/10.5281/zenodo.24420522; repository https://github.com/NeoCortex0815/knowing-is-not-committing.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. *arXiv preprint arXiv:2412.06464*, 2024.

Zejia You, Chunyuan Deng, and Hanjie Chen. Spherical steering: Geometry-aware activation rotation for language models. In *International Conference on Machine Learning (ICML)*, 2026.

Igor M. Zhuk. Weight-space audit of MIT-p50c and DNA weight editing (SGNE-CC) on Qwen3.5-0.8B at a 32,768-token generation limit, 2026a. Zenodo archive, https://doi.org/10.5281/zenodo.24235218; code https://github.com/Nondual-Observer/dna-weight-editing.

Igor M. Zhuk. SGNE-CC v2: Editing 205 weight columns in DeepSeek-R1-Distill-Qwen-1.5B improves greedy accuracy and reduces unfinished responses, 2026b. Zenodo archive, https://doi.org/10.5281/zenodo.24398165; code https://github.com/Nondual-Observer/deepseek-r1-1.5b-sgne-cc-v2.

A Additional Empirical Breakdowns and Benchmark Cards

A.1 Alternative MATH-500 Grading Rules and Multilingual MGSM-240 Breakdown (Study I)

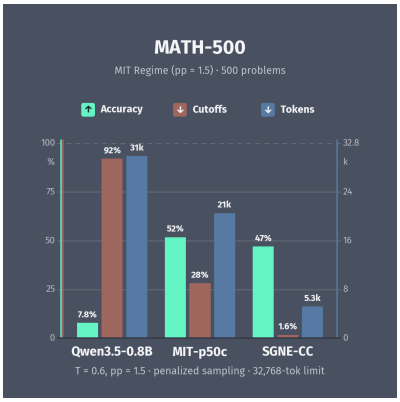
Table 5 compares strict grading against the MIT reference primary and lenient fallback extractors on all 500 MATH-500 problems under pure greedy decoding ($T = 0$, no presence penalty). Table 6 reports the per-language greedy breakdown across the six MGSM-240 languages (40 parallel problems per language), and Figures 3–4 present all remaining benchmark cards across both studies (17 cards in total across Figures 1–4).

Table 5: **MATH-500 sensitivity to symbolic answer-extraction rules** ($N = 500$, Study I, $T = 0$, no presence penalty). Accuracy is in percentage (%).

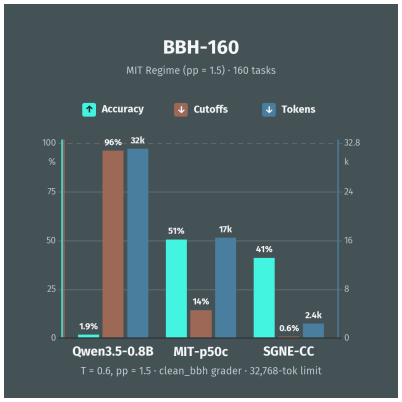
Symbolic Grading Rule	Base (↑)	MIT-p50c (↑)	SGNE-CC v1 (↑)	v1 vs. MIT (McNemar)
Strict (closed `</think>`, EOS, boxed answer)	0.0 (0)	5.0 (25)	**43.0** (215)	$+195/-5$, 3.24×10^{-51}
MIT primary (symbolic check without requiring EOS)	2.8 (14)	6.4 (32)	**43.0** (215)	$+188/-5$
MIT fallback (extracting from truncated 32k drafts)	14.2 (71)	18.6 (93)	**44.2** (221)	$> 2.3\times$ solved ratio

B Offline Reproducibility Protocol and SHA-256 Digests

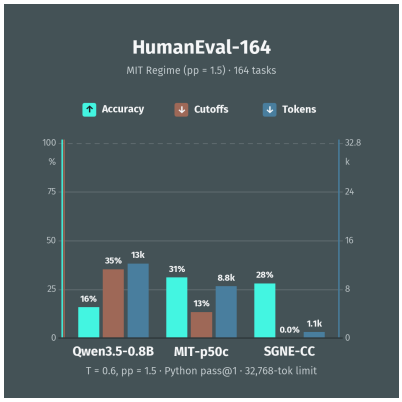
Public repositories, BF16 patches, and archival Zenodo releases for Study I (https://github.com/Nondual-Observer/dna-weight-editing, https://doi.org/10.5281/zenodo.24235218, Hugging Face `Nondual-Observer/Qwen3.5-0.8B-SGNE-CC`) and Study II (https://github.com/Nondual-Observer/deepseek-r1-1.5b-sgne-cc-v2, https://doi.org/10.5281/zenodo.24398165, Hugging Face



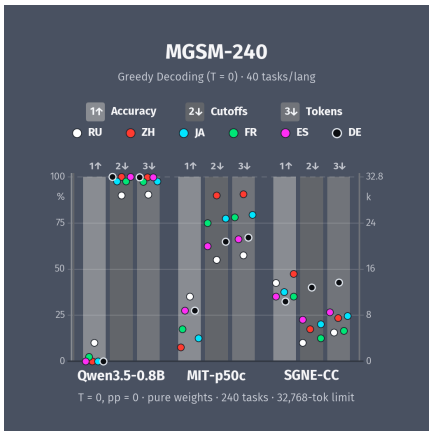
(a) **MATH-500** ($T = 0.6$, **pen.** 1.5)



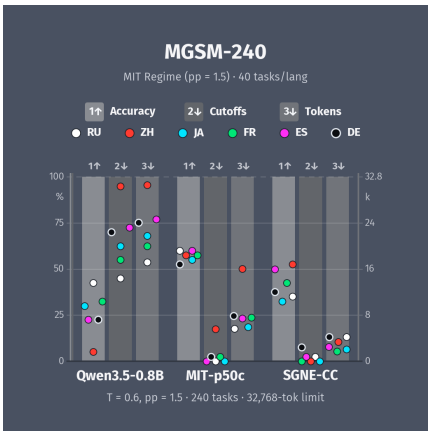
(b) **BBH-160** ($T = 0.6$, **pen.** 1.5)



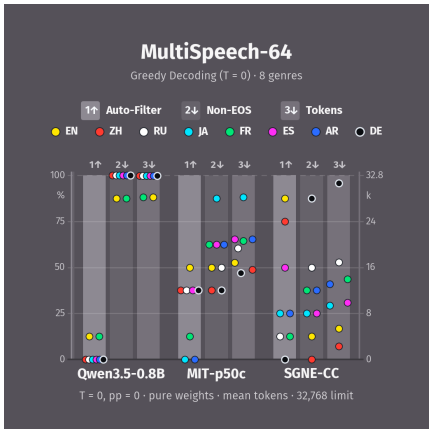
(c) **HumanEval-164** ($T = 0.6$, **pen.** 1.5)



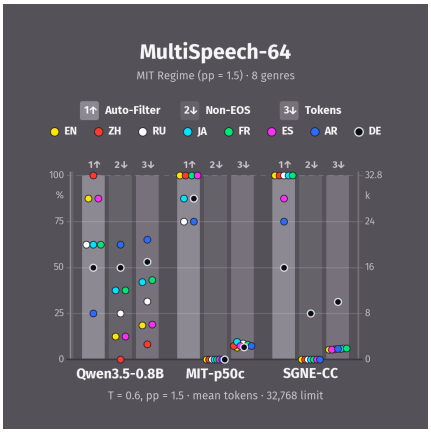
(d) **MGSM-240 (Greedy,** $T = 0$)



(e) **MGSM-240** ($T = 0.6$, **pen.** 1.5)



(f) **MultiSpeech-64 (Greedy,** $T = 0$)

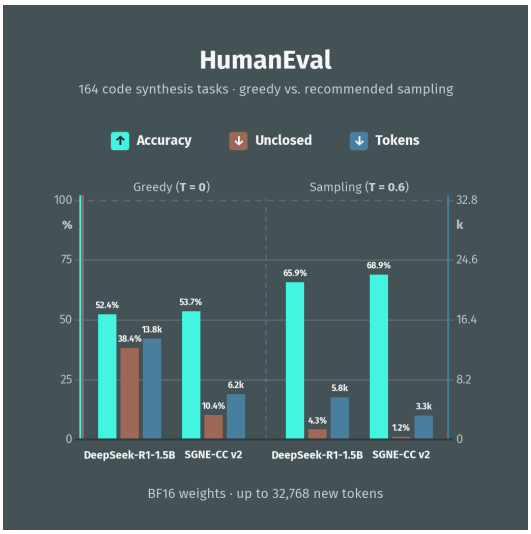


(g) **MultiSpeech-64** ($T = 0.6$, **pen.** 1.5)

Figure 3: **Presence-penalized sampling and multilingual distributions in Study I on Qwen3.5-0.8B.** On the cards, `SGNE-CC` denotes SGNE-CC v1, `MIT Regime` denotes the MIT-p50c sampling setup ($T = 0.6$, presence penalty 1.5), and footer `pp` denotes presence penalty. Panels (a–c) show grouped bar charts under penalized sampling; panels (d–e) plot per-language circular markers across six languages on MGSM-240 (40 tasks/language), and panels (f–g) plot eight languages on MultiSpeech-64 (8 genres/language), where vertical lanes `1↑` (`Accuracy` / `Auto-Filter`) and `2↓` (`Cutoffs` / `Non-EOS`) use the left percentage axis (%) and lane `3↓` (`Tokens`) uses the right axis (k).

Table 6: **MGSM-240 per-language greedy breakdown** (40 problems per language, Study I, $T = 0$, no presence penalty). Accuracy and cutoffs are in percentage (%).

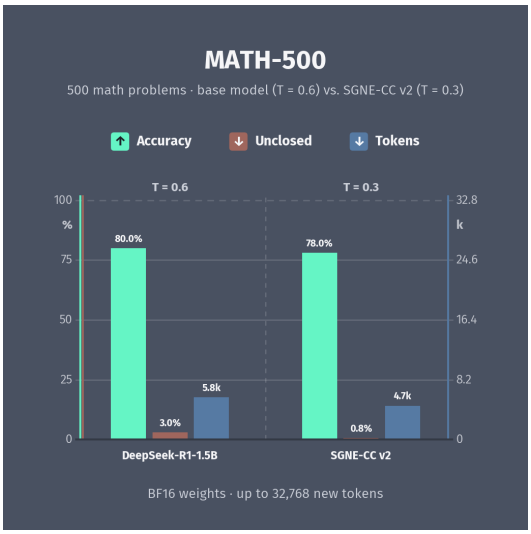
Language (40 items)	**Accuracy (%, ↑)**			**Cutoffs (%, ↓)**		**Mean Tokens (↓)**	
	Base	**MIT-p50c**	**SGNE-CC v1**	**MIT-p50c**	**SGNE-CC v1**	**MIT-p50c**	**SGNE-CC v1**
Russian (RU)	10.0	35.0	**42.5**	55.0	**10.0**	18.9k	**5.1**k
Chinese (ZH)	0.0	7.5	**47.5**	90.0	**17.5**	29.8k	**7.7**k
Japanese (JA)	0.0	12.5	**37.5**	77.5	**20.0**	26.0k	**8.1**k
French (FR)	2.5	17.5	**35.0**	75.0	**12.5**	25.6k	**5.5**k
Spanish (ES)	0.0	27.5	**35.0**	62.5	**22.5**	21.7k	**8.7**k
German (DE)	0.0	27.5	**32.5**	65.0	**40.0**	22.0k	**14.0**k



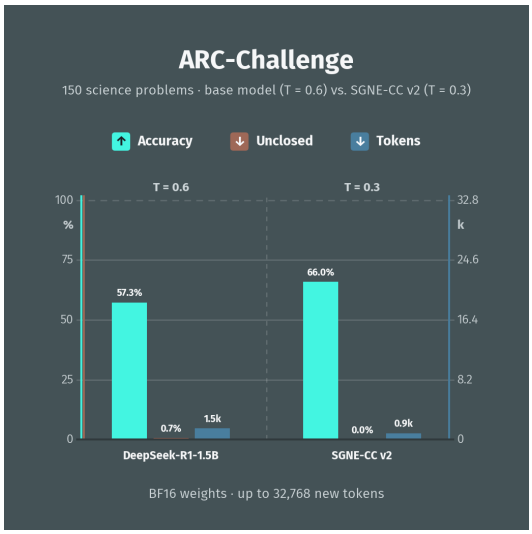
(a) HumanEval-164 ($T = 0$ **vs.** 0.6)



(b) ARC-Challenge-150 ($T = 0$ **vs.** 0.6)



(c) MATH-500 (Base $T = 0.6$ **vs. v2** $T = 0.3$**)**



(d) ARC-Chall.-150 (Base $T = 0.6$ **vs. v2** $T = 0.3$**)**

Figure 4: **Out-of-domain code and science benchmarks and $T = 0.3$ operating-point checks in Study II.** Evaluated on `DeepSeek-R1-Distill-Qwen-1.5B` with no presence penalty at a 32,768-token budget (left axis: % for `Accuracy` ↑ and `Unclosed` ↓; right axis: k for `Tokens` ↓): (a) HumanEval-164 and (b) ARC-Challenge-150 at matched temperatures ($T = 0$ and $T = 0.6$); (c) MATH-500 and (d) ARC-Challenge-150 comparing Base at $T = 0.6$ against SGNE-CC v2 at $T = 0.3$.

Nondual-Observer/DeepSeek-R1-Distill-Qwen-1.5B-SGNE-CC-v2) provide full scripts and trace artifacts for offline CPU verification of weight-space geometry and per-item metric recomputation without GPU access and without executing untrusted model-generated code by default. For dataset rights compliance on MATH-500, both verbatim problem prompts and raw <think> prose scratchpads (which can quote problem statements) are replaced in both studies by unique_id keys, extracted final answers, completion/token/repetition metrics, and SHA-256 digests, while full raw responses are retained on MGSM-240, BBH, ARC-Challenge, HumanEval, MBPP, and MultiSpeech-64:

- **Study I (Qwen/Qwen3.5-0.8B):**
  - Base model.safetensors SHA-256: 04b1c301231dd422b8860db31311ab2721511346a32cb1e079c4c4e5f1fe4696
  - Weight patch sgne_cc_dna_delta.npz (3.13 MiB) SHA-256: 4e4d36e032869bc10f56e2a2033a52d3f2774321cf867ef6f74274a7c0d1566d
  - Holdout and suite manifest suites/reproducibility_and_holdout_manifest.json: c36bd257d55e7643f0c10378fdd935f184e1740630b7c0a05f42f48cea01689a
  - Verification commands: run python3 scripts/verify_paper_metrics.py to verify all 52 SHA-256 digests, run symbolic math-verify equivalence on all extracted answers, check ast.parse syntax and archived unit-test verdicts on HumanEval-164, and recompute McNemar and 40-block sign-flip tests; run python3 apply_dna_patch.py --base-model /path/to/qwen --delta sgne_cc_dna_delta.npz --verify-only (or --compare with MIT-p50c) to audit FP64 column norms and singular modes.
- **Study II (deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B):**
  - Base model.safetensors SHA-256: 58858233513d76b8703e72eed6ce16807b523328188e13329257fb9594462945
  - Weight patch patch/FINAL_SGNE_CC_V2_205_PATCH.npz (0.58 MiB) SHA-256: 6f435a63eb306047fe55353b7a911d32a17dd999efa65f082dd1d7c3779a0b2c
  - Manifest and per-item certificate (verification/SHA256_AND_METRICS_CERTIFICATE.json and verification/per_item_outcomes_and_exclusions.json; 43/43 SHA-256 digests across 5,270 evaluation records, including explicit ID lists for $N = 474, 421, 406, 392$).
  - Verification commands: run python3 verification/verify_repo_integrity.py (add --execute-code-tests inside an isolated sandbox to re-run all HumanEval/MBPP unit tests live); run python3 patch/apply_sgne_cc_v2_205_to_safetensors.py --base-dir /path/to/base --patch patch/FINAL_SGNE_CC_V2_205_PATCH.npz --out-dir /path/to/output to assemble the edited BF16 checkpoint.

Authorship and AI collaboration. This study was conceived, designed, and verified by Igor M. Zhuk in collaboration with AI systems (Gemini, Claude, and OpenAI Codex), which assisted with GPU evaluation harnesses, SHA-256 verification scripts, figures, and English LaTeX drafting.