MOIR: Let the Model Direct Its Own Story for Robust Cross-Domain Knowledge Editing

Jea Kwon · Jiwon Kim · Dong-Kyum Kim · Meeyoung Cha

Max Planck Institute for Security and Privacy (MPI-SP), Bochum, Germany

What should knowledge editing preserve? Let the model answer.

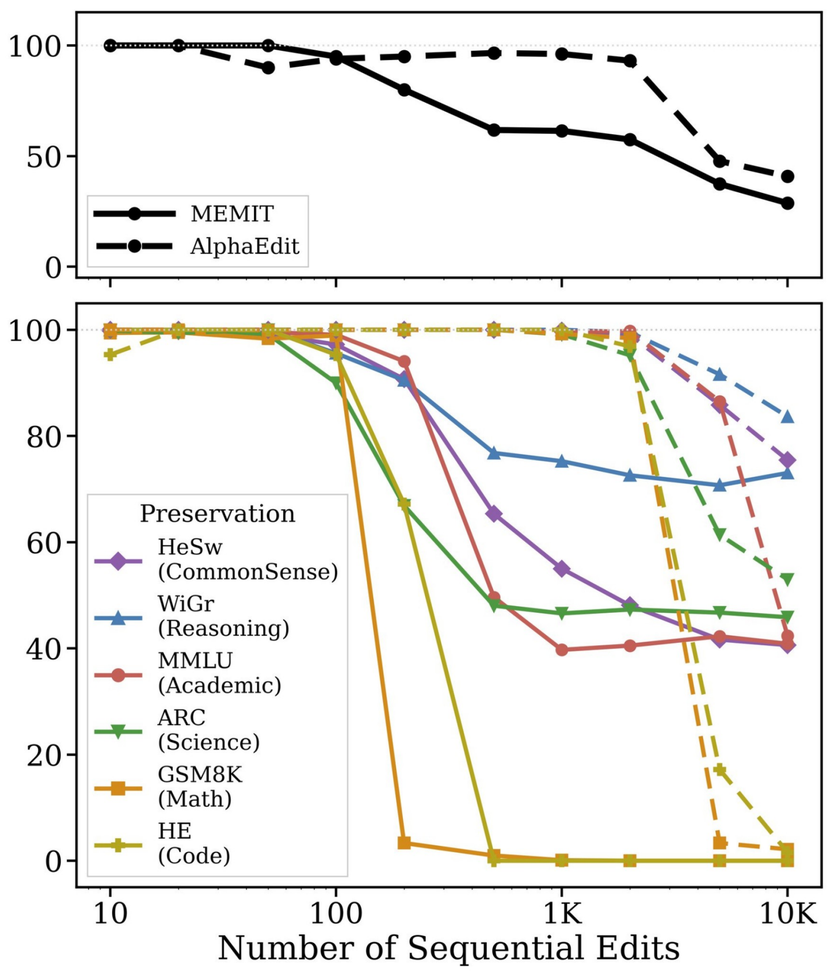
Covariance-constrained editors protect only the subspaces their reference corpus spans. **MOIR** estimates the preservation covariance C from the model's own generations, seeded by a single random token, so preservation aligns with the operative distribution that post-training actually produced. Data-free, drop-in for MEMIT and AlphaEdit, no retraining.

GSM8K ACCURACY · 20,000 ALPHAEDIT EDITS · QWEN3-8B

79.9% with C_{Moir} (self-generated) *vs* **10.9%** with C_{Wiki} (Wikipedia)

1 Editing collapses math & code

Locate-then-edit methods constrain each update with a preservation covariance $C = K_0 K_0^T$, conventionally estimated from Wikipedia. What this proxy misses is sharply asymmetric: encyclopedic recall (MMLU, WinoGrande) survives thousands of edits, while math and code (GSM8K, HumanEval) collapse to near zero.

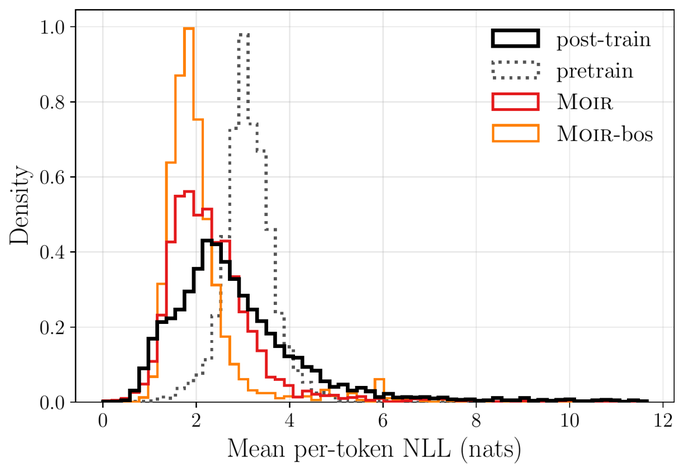


Sequential MEMIT edits on OLMo-2-7B-Instruct with Wikipedia-based C: GSM8K and HumanEval collapse orders of magnitude earlier than encyclopedic tasks.

2 Why: *C* decides what survives

Proposition 1 (G-independence). For closed-form covariance-constrained editors, the update ΔW depends on the input distribution only through C (the K-FAC input factor); the output-side factor G cancels in the closed form. Choosing the corpus is choosing the geometry.

Post-training (SFT, DPO) shifts the operative manifold away from the raw pretraining mixture. Wikipedia stays anchored to pretraining, and even an oracle C built from the true pretraining mix is matched by self-generation:



After 2K MEMIT batch edits (OLMo-2-7B-Inst):

C source	GSM8K	HE
Wikipedia	35.1%	14.0%
OLMo-Mix (oracle)	48.5%	23.7%
MOIR (self-gen.)	**48.3%**	**26.8%**

The true pretraining mixture helps, but self-generation matches the oracle with no data access at all.

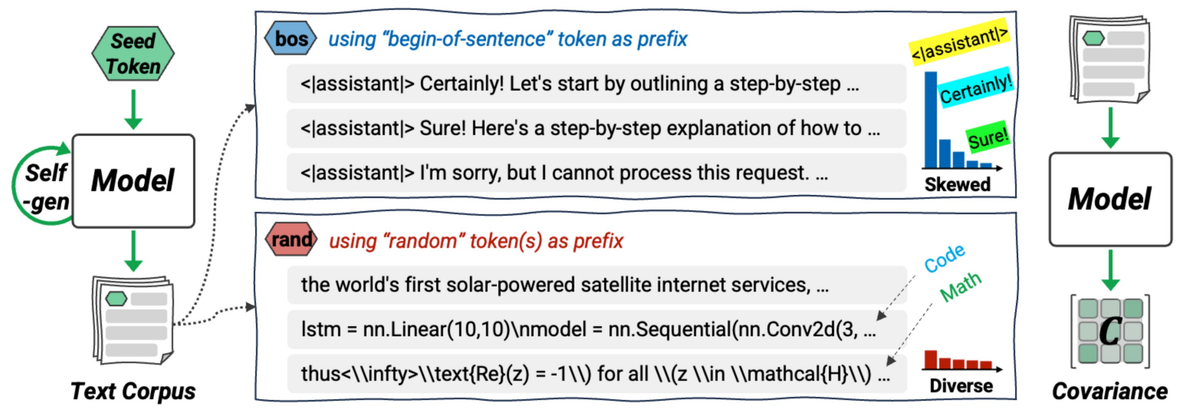
Per-document NLL under OLMo-2-7B-Instruct: Moir (rand × 1) tracks the shifted post-training distribution; bos-only and static corpora stay misaligned (JSD 3–4× larger).

Limitations

Two locate-then-edit families tested (MEMIT, AlphaEdit); the findings should extend to any covariance-like preservation operator, but this remains to be verified. Extreme RLHF-style mode collapse could under-sample rare capabilities. Experiments cover 7–8B instruction-tuned models; frontier scale and base models are future work.

3 MOIR: estimate *C* from the model itself

MOIR (**M**emories **O**f **I**nternal **R**epresentations) samples 100K sequences from the deployed model's own decoding distribution, collects the MLP keys $\boldsymbol{k}$ during the forward pass, and sets $C = (1/N) \Sigma \boldsymbol{k}\boldsymbol{k}^T$. No external data, no architecture changes: a drop-in C for MEMIT, AlphaEdit, or any covariance-based editor.



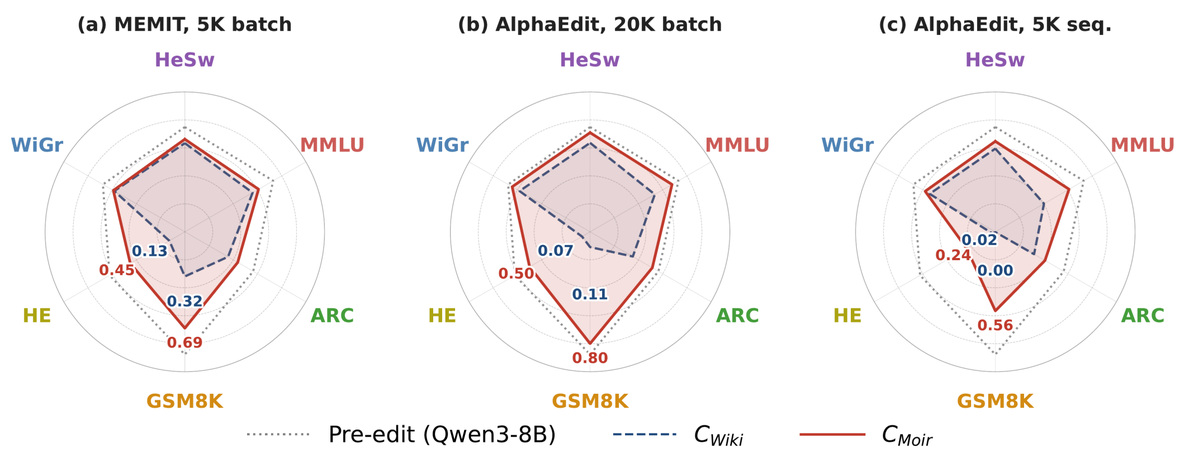
Seeding with bos collapses onto instruction-style openings; a single random token diversifies generation across the model's internalized domains (text, code, math).

52.8% of `<bos>`-seeded samples open with `<|assistant|>`: mode collapse

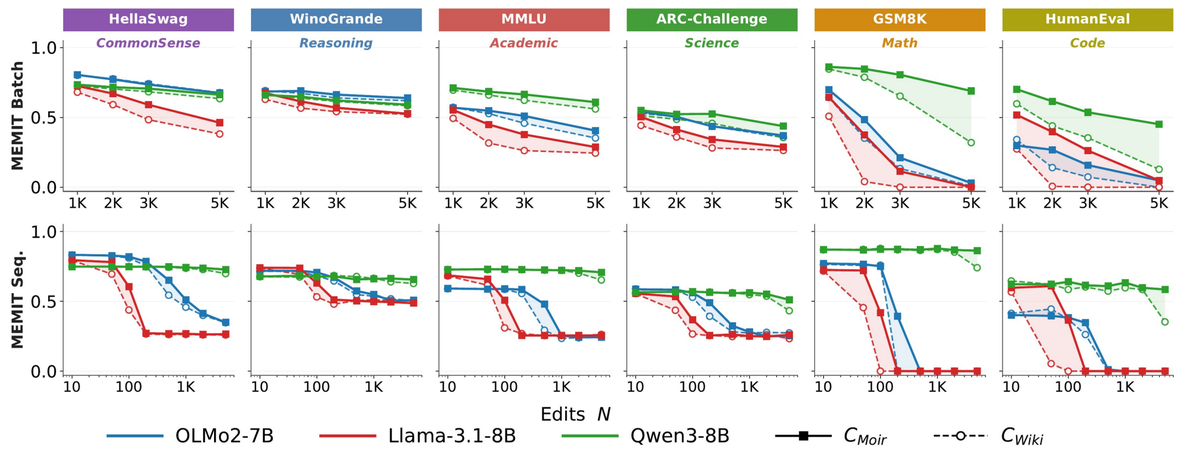
0.66% max first-token frequency with `<rand>×1`: attractor broken

k = 1 single-token seeding is the sweet spot: total score 1.093 vs 1.006 for `<bos>`

4 Results: collapse averted



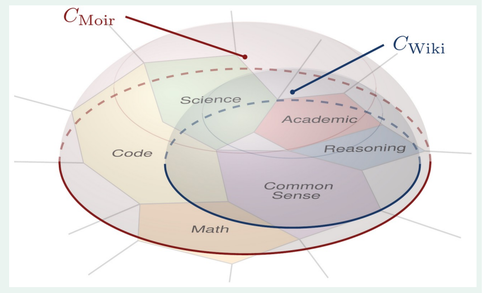
Qwen3-8B after editing: Wikipedia-C keeps encyclopedic axes while GSM8K / HumanEval collapse; the self-generated C restores them without disturbing what was already preserved.



Per-task preservation under MEMIT (batch & sequential, three models): shaded bands mark the Moir–Wikipedia gap; editing quality is matched or improved throughout.

5 Takeaways

- What to preserve matters as much as how: the corpus behind C decides which capabilities survive an edit.
- Post-training moves the model off every static corpus, including its own pretraining mix; the live model is the most accessible source of its operative distribution.
- One-time cost: ~2 h of self-generation per model; covariance estimation ≈ 15 min on one A100.
- Once edits stop costing reasoning and code, editing becomes maintainable infrastructure for keeping deployed LLMs current.



C_{Moir} covers the regions the model actually uses; C_{Wiki} spans only the encyclopedic chart.

Models: OLMo-2-7B-Instruct · Llama-3.1-8B-Instruct · Qwen3-8B | Editors: MEMIT & AlphaEdit (batch + sequential) | Preservation suite: MMLU · GSM8K · HellaSwag · WinoGrande · ARC-Challenge · HumanEval