

21 OCTOBER 2026 / INTERIM FINDINGS / v1.1

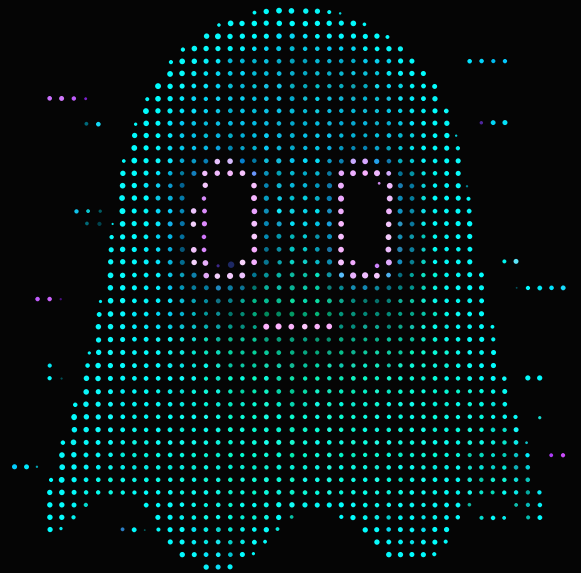
GH0ST

Weight-only quantization and Solidity execution *for sub-billion language models.*

An empirical study of numerical quality,
weight storage and matrix-kernel gas.

GH0ST RESEARCH

Twenty-checkpoint campaign.
Ten completed searches in this snapshot.
Evidence pinned to immutable source commits.



COVER / NEURAL APPARITION

SNAPSHOT 21 OCT 2026 / 01:28:06 UTC

RECORD 715 CANDIDATES / 8 FINALIZED GAS REPORTS

FIRST RESEARCH PAPER / GH0ST RESEARCH SERIES

Abstract

Interim results from a twenty-checkpoint campaign.

We examine the storage, numerical-quality and execution-gas constraints of representing sub-billion-parameter language models in Solidity. A twenty-checkpoint campaign assigns at least twenty minutes of active quantization search to each model. This interim snapshot contains ten completed searches, 715 candidate configurations and eight finalized gas reports. Weight-only affine quantization uses four-bit codes, integer scaling and optional FP32 embedding protection; candidates are selected using a separate validation partition. Nine of ten selected candidates increase perplexity on the small held-out diagnostic corpus. Selected payloads require an estimated 1,728-8,644 Robinhood data shards. Across 72 matrix-tile/input cases, 432 local EVM execution receipts match an independent integer reference. Comparing baseline and optimized compiler configurations reduces random-input tile gas by 38.43-42.97%; matched-run kernel changes alone save 0.48-0.52%. These results identify execution trade-offs and evaluation needs; they do not establish preserved general capability, complete on-chain inference or an affordable public chat service.

KEYWORDS / weight quantization; EVM; Solidity; gas; language models

10

COMPLETED SEARCHES

715

CANDIDATES TESTED

8

FINALIZED GAS REPORTS

Research question

How far can weight-only Q4 compression and Solidity matrix execution reduce the storage and compute burden of a larger GHOST model while retaining useful numerical quality?

Three endpoints, reported separately

The study measures selected full-model quantization, accounted storage, and separately prepared matrix-kernel execution. These endpoints share checkpoints, but a gas result does not by itself validate the selected model representation.

READING GUIDE	PAGE	READING GUIDE	PAGE
Study design & methods	03	Discussion & limitations	07
Quality & precision	04	Campaign register	08
Storage & deployment	05	Reproducibility	09
Matrix gas & compilers	06	Pinned source ledger	10

Study design & methods

Fixed search budgets and held-out diagnostics.

1. Scope and cohort

The cohort contains twenty open checkpoints below one billion unique parameters, spanning Qwen, LiquidAI LFM2, Hugging Face SmoLLM and EleutherAI Pythia. Base and instruction-tuned checkpoints count separately. The selection is a varied research shortlist; it is not a validated ranking of the twenty best models. At the frozen timestamp, ten searches were complete [R01-R10].

2.1 Active optimization budget

Each completed search ran for at least 1,200 seconds of elapsed active search. Model download, calibration, baseline evaluation, export and final checks were outside that timer. Candidate counts therefore vary with model cost. The ten completed searches tested 715 configurations in total. Branch checkpoints are published every 600 seconds.

2.2 Weight representation

Eligible two-dimensional weight tensors use affine four-bit codes in groups of 64 or 128. Search choices include clipping ratios, zero-point offsets, diagonal calibration weighting (power 0 or 1), and optional FP32 protection of embedding/output parameters. Nonmatrix tensors remain FP32. These are mixed-precision representations.

QUANTIZATION RECORD

Reconstruction: $w_{\text{hat}} = (q - z) \times s \times u / 2^{24}$.

Each weight stores a four-bit code q ; each group stores a four-bit zero point z and a twelve-bit scale s . The shared tensor unit is u . Metadata uses two bytes per group, plus a sixteen-byte header per quantized matrix.

2.3 Selection and evaluation

Select the smallest accounted payload whose validation perplexity is no more than 15% above baseline. If no candidate passes, retain the lowest-validation-loss candidate and flag the gate as failed. The held-out test does not select candidates. Perplexity is $\exp(\text{token-weighted next-token cross entropy})$; the first token in each document is excluded from scoring.

PARTITION	DOCUMENTS	SCORED TOKENS
Calibration	8	Not logged
Validation	4	Qwen 72 / LFM2 76 / SmoLLM2 71
Held-out test	8	Qwen & SmoLLM2 140 / LFM2 152
Generation probes	4 prompts	Greedy / at most 24 new tokens

All text partitions are disjoint original diagnostic passages. The sample is small and is not a standard capability benchmark. CPU activations remain FP32; no AWQ or GPTQ implementation is claimed.

03 / GH0ST RESEARCH PAPER 001

Quality & selected precision

Validation selection and held-out perplexity.

CHECKPOINT	Q4 GROUP	FP32 E/H*	BASE PPL	TEST PPL	CHANGE	VAL. GATE
Qwen3-0.6B R01	128	No	50.41	62.33	+23.6%	Pass
Qwen3-0.6B-Base R02	128	Yes	31.98	45.73	+43.0%	Fail
LFM2-700M R03	128	No	54.57	52.13	-4.5%	Pass
LFM2-350M R04	128	No	46.50	50.62	+8.9%	Pass
Qwen2.5-0.5B-Instruct R05	64	Yes	38.96	49.40	+26.8%	Fail
Qwen2.5-Coder-0.5B-Instruct R06	128	No	49.67	62.63	+26.1%	Pass
Qwen2.5-0.5B R07	64	No	33.14	46.22	+39.5%	Fail
Qwen2.5-Coder-0.5B R08	128	No	33.39	64.29	+92.5%	Pass
SmoLM2-360M-Instruct R09	64	No	29.67	39.89	+34.4%	Pass
SmoLM2-135M-Instruct R10	64	Yes	47.43	54.41	+14.7%	Fail

Table 1. Selected candidates for completed searches [R01-R10]. E/H = embedding/output protection. Gas finalization was pending for R09-R10; their quantization and text results were complete. Lower PPL is better within each checkpoint.

TEST PPL CHANGE RELATIVE TO EACH CHECKPOINT BASELINE

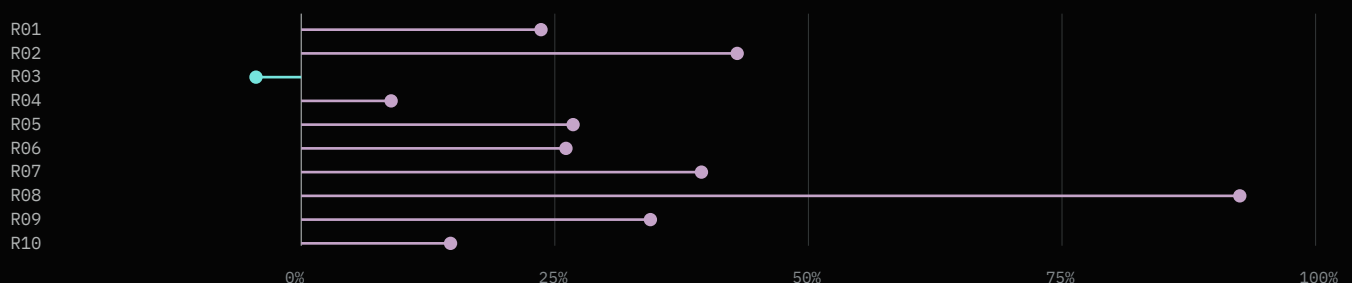


Figure 1. Relative test-PPL change for each selected candidate, using the baseline for that same checkpoint. Raw PPL across tokenizers is not a model leaderboard.

Validation does not guarantee held-out retention

Six selections passed validation; nine increased test PPL. LFM2-700M alone improved (54.57 to 52.13).

Qwen2.5-Coder-0.5B passed validation but test PPL rose 92.5%. With only 140-152 scored test tokens, these diagnostic observations do not establish general capability or statistical significance.

The storage boundary

Mixed-precision payloads and deployment lower bounds.

SELECTED CHECKPOINT	PARAMS M	PAYLOAD MB	RH SHARDS	DEPOSIT B GAS
Qwen3-0.6B	751.63	387.80	3,945	77.56
Qwen3-0.6B-Base	596.05	849.69	8,644	169.94
LFM2-700M	742.49	383.19	3,899	76.64
LFM2-350M	354.48	183.01	1,862	36.60
Qwen2.5-0.5B-Instruct	494.03	734.93	7,477	146.99
Qwen2.5-Coder-0.5B-Instruct	494.03	255.00	2,594	51.00
Qwen2.5-0.5B	494.03	262.72	2,673	52.54
Qwen2.5-Coder-0.5B	494.03	255.00	2,594	51.00
SmoLM2-360M-Instruct	361.82	192.45	1,958	38.49
SmoLM2-135M-Instruct	134.52	169.80	1,728	33.96

Table 2. Unique loaded parameter counts, selected payload accounting and deployment lower bounds [R01-R10]. MB = 1,000,000 bytes; B gas = 1,000,000,000 gas. Counts use loaded parameters rather than nominal checkpoint names.

4.1 Accounting assumptions

Whole-model accounting includes a sixteen-byte container header and a thirty-two-byte descriptor per tensor. Each retained FP32 parameter costs four bytes. Quantized groups and matrices use the format described in Section 2.2. The RH shard estimate assumes 98,303 payload bytes per data contract; the standard EVM comparison uses 24,575 bytes. These limits are accounting assumptions, not new mainnet probes.

CODE-DEPOSIT LOWER BOUND

$N = \text{ceil}(\text{payload bytes} / 98,303)$

Code-deposit gas = $200 \times (\text{payload bytes} + N)$

The extra byte per shard is a STOP byte. Full creation overhead, indexes, execution contracts, calldata, parent-chain data fees and gas price are excluded.

4.2 Consequence for deployment

Selected outputs span 169.80-849.69 MB and an estimated 1,728-8,644 RH data shards. None meets a 500-contract target. FP32 embedding/output protection can dominate the payload: Qwen3-0.6B-Base selected 849.69 MB and still failed the validation gate. A smaller nominal model is therefore not automatically the smallest selected representation.

These figures are storage estimates and gas lower bounds, not complete deployment budgets or ETH fee quotes.

Matrix gas & compiler trade-offs

Execution receipts under three build configurations.

All values below are local Cancun EVM receipt gas for eight output rows with deterministic random int16 input. Tile data and inputs are identical within each comparison. Shared shapes are consolidated; source tensors and other input patterns remain in the linked reports [R01-R08]. No public transactions were broadcast by this campaign.

TILE	BASE 200	AFFINE 200	AFFINE 200k	MATCHED*	COMBINED*
8 × 896	233,335	232,210	143,666	0.48%	38.43%
8 × 1,024	262,452	261,174	160,000	0.49%	39.04%
8 × 1,536	379,043	377,153	225,456	0.50%	40.52%
8 × 2,048	495,692	493,190	290,971	0.50%	41.30%
8 × 3,072	729,033	725,307	422,043	0.51%	42.11%
8 × 4,608	1,079,786	1,074,224	619,392	0.52%	42.64%
8 × 4,864	1,138,306	1,132,438	652,345	0.52%	42.69%
8 × 6,912	1,606,967	1,598,651	916,469	0.52%	42.97%

Table 3. Baseline128 at optimizer runs 200; Affine128 at runs 200 and 200,000. *Matched compares the two runs-200 builds. Combined compares Baseline128/200 with Affine128/200,000.

ATTRIBUTION MATTERS

The matched-run implementation change saves 0.48-0.52% on these random-input tiles. The 38.43-42.97% saving includes both the implementation change and a higher compiler optimizer setting. The experiment does not isolate a baseline build at 200,000 runs.

5.1 Execution validation

Seventy-two tile/input cases cover random, minimum, maximum and alternating input vectors across eight reports. Three builds ran twice per case, producing 432 execution receipts with exact agreement against an independent integer reference. Repeated receipts are deterministic checks, not independent statistical samples.

5.2 One-time deployment trade-off

BUILD / OPTIMIZER RUNS	RUNTIME BYTES	DEPLOYMENT GAS
Baseline128 / 200	4,429	1,008,686
Affine128 / 200	6,514	1,458,483
Affine128 / 200,000	8,469	1,862,876

Table 4. One-time local kernel deployments, excluding weight shards. Higher-run Affine128 has lower tested execution gas but larger deployed code. These measurements do not include complete inference.

Discussion & limits

Interpretation, unresolved questions and next experiments.

6.1 What this study establishes

The campaign provides a reproducible quantization search record, explicit payload accounting and exact-reference checks for a shared Solidity matrix primitive. The compiler configuration materially changes execution gas. The available quality results also show why storage selection needs evaluation beyond a short validation set. These are distinct engineering findings.

6.2 What remains unresolved

The gas tiles are separately requantized from original checkpoint weights into a fixed affine128 format with clipping ratio 1 and importance power 0. Their settings can differ from the selected full-model configuration. Group64 selections have no matching group64 execution-kernel validation in this snapshot. Tile agreement therefore cannot be read as full-model fixed-point agreement.

Attention, recurrent state, normalization, KV caches, generation control, scheduling and full answer production are outside these tile measurements. This cohort has no validated end-to-end on-chain model runtime or measured public-chain inference fee. The report adds no Stylus benchmark and makes no RH-mainnet inference-cost claim.

6.3 Threats to validity

The hand-written diagnostic corpus contains only eight test passages, with 140-152 scored tokens per checkpoint. The search uses a short validation partition, so selection may fit that partition without generalizing. No task-accuracy, reasoning, coding, multilingual or autonomous-agent benchmark is established. Perplexity values from different tokenizers are not directly comparable.

The twenty-minute wall-time budget is consistent, but candidate counts vary with architecture and execution cost. CPU model and memory capacity are not recorded. The campaign is incomplete, the shared kernels are not a comprehensive optimizer sweep, and the reported gas is local execution under one EVM configuration.

6.4 Next experiments

Expand held-out text and task evaluation; compare quantization settings on untouched test data; validate execution for the selected representation, including group64; add matched compiler sweeps for both kernel families; then measure a complete fixed-point generation path and its deployment requirements.

CONCLUSION

The current evidence supports a documented research path toward larger on-chain models. Practical deployment still requires stronger quality retention, substantially smaller storage and validated end-to-end execution.

07 / GH0ST RESEARCH PAPER 001

Campaign register

Per-checkpoint progress at the frozen snapshot.

#	CHECKPOINT	STATE	TRIALS	MINUTES	SOURCE
01	Qwen3-0.6B	Finalized	39	20.38	R01
02	Qwen3-0.6B-Base	Finalized	44	20.04	R02
03	LFM2-700M	Finalized	33	20.02	R03
04	LFM2-350M	Finalized	69	20.33	R04
05	Qwen2.5-0.5B-Instruct	Finalized	32	20.14	R05
06	Qwen2.5-Coder-0.5B-Instruct	Finalized	55	20.26	R06
07	Qwen2.5-0.5B	Finalized	87	20.09	R07
08	Qwen2.5-Coder-0.5B	Finalized	61	20.08	R08
09	SmolLM2-360M-Instruct	Gas pending	70	20.22	R09
10	SmolLM2-135M-Instruct	Gas pending	225	20.02	R10
11	SmolLM2-360M	Preparing	-	-	R11
12	SmolLM2-135M	Queued	-	-	R12
13	SmolLM-360M	Queued	-	-	R13
14	SmolLM-135M	Queued	-	-	R14
15	SmolLM-360M-Instruct	Queued	-	-	R15
16	SmolLM-135M-Instruct	Queued	-	-	R16
17	Qwen2-0.5B-Instruct	Queued	-	-	R17
18	Qwen2-0.5B	Queued	-	-	R18
19	pythia-410m	Queued	-	-	R19
20	pythia-160m	Queued	-	-	R20

Table 5. Frozen status at 01:28:06 UTC on 21 October 2026. Twenty model branches exist; ten searches completed. Eight reports were finalized, two awaited gas finalization, one model was preparing and nine were queued.

Recovery without repeating the search

The original environment lacked pycryptodome after optimization. The first finalizer recovered four reports, then encountered transient GitHub REST disconnections. A revised finalizer uses authenticated Git transport with bounded retries. Eight reports were finalized by this snapshot; original failures, selected manifests, candidate measurements and text results remain preserved.

A finalized report means its gas checks completed. It does not mean the selected quantization passed the quality gate or that the checkpoint is ready for deployment.

Reproducibility & disclosures

Environment, evidence and publication status.

8.1 Recorded environment

COMPONENT	RECORDED VALUE
Python / GCC	3.12.15 / 13.3.0
PyTorch / execution	2.10.0+cpu / 2 threads
Transformers / NumPy	5.2.0 / 2.2.6
Solidity compiler	0.8.30+commit.73712a01.Emscripten.clang
Platform	Linux 7.0.0-1017-azure / x86_64 / glibc 2.39
EVM benchmark target	Local Cancun
CPU model / RAM	Not recorded

The campaign README declares Node 22 and Anvil 1.8.5. These declared toolchain versions are distinct from values recorded in each result environment. No unrecorded hardware characteristics are inferred.

8.2 Immutable evidence

Appendix A pins the exact report commits used for this edition. The linked JSON files preserve full upstream revisions and, where produced, checkpoint SHA-256 records, candidate settings, per-document losses, generation probes, environment details and gas receipts. This edition retains the original snapshot; later branch progress is not silently mixed into these results.

8.3 Research record

[C1] Campaign code, protocol and model manifest

[C2] Original sequential optimization workflow

[C3] Gas finalization workflow with Git retries

[C4] GHOST Observatory and project record

8.4 Attribution and publication status

Prepared by GHOST Research as Research Paper 001, an interim technical report. Original checkpoint authors retain attribution and their model licenses continue to apply. This edition does not claim peer review, a registered DOI, benchmark certification or established world-first priority.

8.5 Editorial note

Version 1.1 redesigns the first progress report in the GHOST Observatory visual system and clarifies optimizer-run attribution. The underlying frozen measurements are unchanged. Main numerical results are generated directly from the same structured snapshot.

A / GHOST RESEARCH PAPER 001

Pinned source ledger

Immutable report commits and upstream model revisions.

Each JSON link resolves to the exact model-branch report commit used here. Full upstream model revisions and file hashes remain inside those reports. Revision prefixes below are sixteen characters; report commits are shown in full.

REF	CHECKPOINT	REPORT COMMIT / FULL SHA	MODEL REVISION	FILE
R01	Qwen3-0.6B	3aba1ac5c5f024d4b63cbbe677485c8234f1e5c4	c1899de289a04d12	JSON
R02	Qwen3-0.6B-Base	4fa5c878b144c11b36d8a7ed3532c1f2569b7c3a	da87bfb608c14b7c	JSON
R03	LFM2-700M	7824222c4161ca41cf673aa12a6641ab7523a8ca	86f49cf9a3800c3a	JSON
R04	LFM2-350M	d42837f06898292b5c81a84061c87d06c7973797	f37d3f5c8c5484bc	JSON
R05	Qwen2.5-0.5B-Instruct	25f959520170fa62b1a363f7d5483b2ffc479b87	7ae557604adf67be	JSON
R06	Qwen2.5-Coder-0.5B-Instruct	4496f55c08dcc6fa23dfaaca0a78258be7744b6	ea3f2471cf1b1f0d	JSON
R07	Qwen2.5-0.5B	95597674751e31a902ee39e9980c656b34c60b84	060db6499f32fa18	JSON
R08	Qwen2.5-Coder-0.5B	be7a2008535d5b1969ce99a2dd0505982144e5e0	8123ea2e9354afb7	JSON
R09	SmoLLM2-360M-Instruct	6f43e9c2f2212b0527e3489b01aef39a363cafc3	a10cc1512eabd3dd	JSON
R10	SmoLLM2-135M-Instruct	10279280327813d1b3847a342a6da1925440a0cb	12fd25f77366fa6b	JSON
R11	SmoLLM2-360M	924bcc7af7e21909348880a5f73c3efa86c817df	f8027fd0eaeaa54c	JSON
R12	SmoLLM2-135M	05ba6c9b1ecf7831ba7c9d96a785adf287b138fe	93efa2f097d58c2a	JSON
R13	SmoLLM-360M	05ba6c9b1ecf7831ba7c9d96a785adf287b138fe	59f7ef243ee09a72	JSON
R14	SmoLLM-135M	05ba6c9b1ecf7831ba7c9d96a785adf287b138fe	1d461723eec654e6	JSON
R15	SmoLLM-360M-Instruct	05ba6c9b1ecf7831ba7c9d96a785adf287b138fe	73b7144f76331266	JSON
R16	SmoLLM-135M-Instruct	05ba6c9b1ecf7831ba7c9d96a785adf287b138fe	fcc320f490e08fdb	JSON
R17	Qwen2-0.5B-Instruct	05ba6c9b1ecf7831ba7c9d96a785adf287b138fe	c540970f9e29518b	JSON
R18	Qwen2-0.5B	05ba6c9b1ecf7831ba7c9d96a785adf287b138fe	91d2aff3f957f99e	JSON
R19	pythia-410m	05ba6c9b1ecf7831ba7c9d96a785adf287b138fe	9879c9b5f8bea905	JSON
R20	pythia-160m	05ba6c9b1ecf7831ba7c9d96a785adf287b138fe	50f5173d932e8e61	JSON

EDITION IDENTITY

GHOST / Research Paper 001 / Interim findings / v1.1

Snapshot: 21 October 2026, 01:28:06 UTC (06:58 IST).

Source collection window: 01:28:04.837-01:28:06.427 UTC.

Twenty branches; ten completed searches; eight finalized gas reports.

Correspondence and open contributions

github.com/8tk8m2kv9f-lang/GHOST

[@lmggh0st / gh0st.wiki](https://lmggh0st.github.io/gh0st.wiki)