Teutonic-I 10B: Decentralized Pretraining through King-of-the-Hill Competition

Jacob Steeves Krzysztof Warchoł Hubert Korzeniewski Antoni Bombała

Bittensor Subnet 3 (Teutonic)

August 2026

Abstract

We report Teutonic-I 10B, a 10-billion-parameter language model produced during a 70-day open competition on Bittensor Teutonic Subnet 3. Rather than synchronously aggregating gradients, independent participants trained challengers to a shared incumbent checkpoint in a sequential "king-of-the-hill" process. A challenger replaced the incumbent only when its reduction in per-sequence cross-entropy loss exceeded a minimum effect size under a conservative bootstrap lower confidence bound. Across 2,163 completed duels, the system accepted 203 new incumbents. The best checkpoint, king #191, obtained an unweighted mean score of 62.28% over 11 reported benchmarks. Under the same aggregation of the reported scores, it exceeded Quasar-Preview 18B (59.78%) and Covenant 72B (57.55%), and achieved the highest score among the compared models on 8 of 11 tasks. These results show that decentralized competition can coordinate a productive search over training data and optimization strategies without prescribing a common training pipeline. They do not, however, isolate the contribution of decentralization from data, compute, or training choices, and the cross-model comparison is limited by reliance on reported benchmark values rather than a fully controlled evaluation. We describe the selection rule, reward mechanism, training-objective changes, empirical trajectory, and limitations of this approach.

Keywords: decentralized training, language model pretraining, king-of-the-hill selection, percentile bootstrap, incentive mechanisms, Bittensor.

1 Introduction

Training frontier language models is usually organized as a centralized optimization process: one organization selects the data mixture, training schedule, infrastructure, and checkpoints. Distributed systems may spread computation across many devices, but generally retain a single optimization objective and synchronously aggregate gradients. An alternative is to decentralize the *search for improvements*. Independent participants can choose their own data, optimization procedures, and hardware, while a shared evaluation mechanism determines which checkpoint becomes the basis for subsequent work.

This paper studies such a mechanism as instantiated by Teutonic Subnet 3 (SN3), a subnet of the Bittensor peer-to-peer intelligence market [12]. The competition fixed the Quasar-10B architecture and a 2,048-token context window, but did not mandate training code, data, duration, or infrastructure. At any time, one checkpoint was designated the *king*. A participant downloaded the current state, trained a challenger, and submitted it for paired evaluation against the king. The challenger was promoted only if its loss improvement was both practically large enough and statistically robust.

The competition ran from June 2 through August 10, 2026. It produced Teutonic-I 10B[1] after 2,163 completed evaluations and 203 successful coronations. The strongest recorded checkpoint appeared on August 1, 60 days after the start. Its reported average across 11 shared downstream benchmarks was 62.28%, compared with 57.55% for Covenant 72B and 59.78% for Quasar-Preview 18B. Parameter count is therefore not sufficient to explain the ordering in this comparison. At the same time, the experiment is observational: it does not establish that the competition mechanism alone caused the performance difference.

Our contributions are: (i) a formal description of sequential checkpoint selection using paired losses and a percentile-bootstrap lower confidence bound; (ii) an account of a reward window that compensates the current and four previous kings; (iii) a quantitative summary of the 70-day training trajectory and the changing global data objective, reconstructed from the public validator record; and (iv) a benchmark comparison and a discussion of the validity, efficiency, and governance questions raised by competitive decentralized pretraining.

[1] Model weights: https://huggingface.co/dendriteholdings/Teutonic-I. Subnet implementation: https://github.com/unarbos/teutonic.

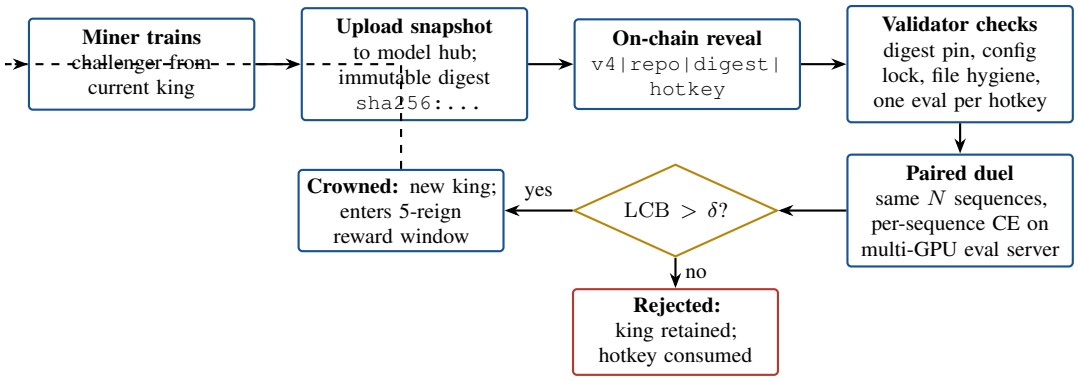

Figure 1: The SN3 king-of-the-hill pipeline. Miners commit an immutable content digest on chain before evaluation, so the validator scores exactly the snapshot that was committed. A successful challenger becomes the new starting point for all subsequent participants (dashed feedback edge).

2 System Design

2.1 Competition protocol

The protocol maintains a single official checkpoint. Let K_t denote the king at competition step t. A miner constructs a challenger C_t with the same architecture, typically beginning from an available checkpoint, and selects its own training procedure. The validator evaluates K_t and C_t on an identical, deterministically selected sample. If C_t passes the statistical criterion in Section 3, then $K_{t+1} = C_t$; otherwise, $K_{t+1} = K_t$. Figure 1 summarizes the full submission and evaluation pipeline.

One way to view the process is as evolution under an extremely strict selection criterion: a new model receives no credit for the effort invested in training it; it must actually outperform the best known version.

This process differs from conventional data- or model-parallel training. Miners do not jointly compute an optimization step, exchange gradients, or average parameters. Coordination occurs through checkpoint selection: every accepted model becomes a new starting point and a public signal about which training choices improved the current objective. The network rewards verified improvement rather than claimed accelerator time, token count, or training effort.

2.2 Fixed and free choices

The competition fixed the model architecture and the 2,048-token context window, the latter dictated by the requirements of the evaluation system. It also defined the evaluation data mixture and acceptance threshold. Miners remained free to choose datasets, preprocessing, curricula, optimizers, hyperparameters, training duration, and hardware. This separation creates a distributed search over training strategies while keeping submitted checkpoints directly comparable.

The setup also creates dependencies. A challenger must outperform the current king, so the value of a training strategy depends on the incumbent and the current evaluation mixture. Improvements are sequential rather than independent, and the reported final model is the product of a lineage of accepted checkpoints.

2.3 Implementation safeguards

Several implementation details of the production validator matter for interpreting the record. First, submissions are pinned to an immutable content digest (an OCI `sha256` manifest digest) committed on chain before evaluation; the validator refuses mutable references, which closes the time-of-check-to-time-of-use gap between commitment and evaluation. Second, a *config lock* requires an exact match of the architecture string and of structural configuration keys (vocabulary, hidden size, layer and head counts, rotary-embedding parameters) against the incumbent; custom modeling code is admitted only if byte-identical to the king's own files. Third, each hotkey registration is granted exactly one evaluation, and each model repository is evaluated at most once, which removes replay and resubmission-until-lucky strategies at the intake layer; a miner who wants another attempt must register (and pay for) a fresh identity. Fourth, the validator refuses to start unless commit-reveal weight setting is enabled on the subnet, as encrypted weight commitments are the load-bearing defense against weight-copying validators. Finally, because copying the public king is the cheapest possible "training" strategy, the validator fingerprints challenger tensors against known checkpoints and compares upload timestamps across hosting backends; a challenger whose weights are identical to an earlier upload is displaced in favor of the earlier submitter without an evaluation. Model-copying attempts were in fact observed during the competition.

Table 1: Parameters of the selection and reward mechanism.

Symbol	Meaning	Value
—	Evaluation sequence length	2,048 tokens
N	Evaluation sequences per duel	global hyperparameter (up to 25,000)
B	Bootstrap replicates	10,000
α	Significance level of the LCB	0.001
δ	Minimum required advantage	0.0015
—	Reward window (kings receiving emissions)	5
—	Incentive share per model in the window	20%

Duels execute on a remote eight-GPU evaluation server that caches the incumbent king across duels, streams progress to the validator over server-sent events, and enforces a hard 30-minute wall-clock budget per evaluation. In the retained public record, the median completed duel took 1,063 seconds (≈17.7 minutes) end to end, during which both models scored up to 25,000 sequences of 2,048 tokens each (≈51M tokens per side).

3 Statistical Checkpoint Selection

3.1 Paired per-sequence loss

The king and challenger are evaluated on the same N sequences, each containing 2,048 tokens. Let L_i^K and L_i^C be their respective mean cross-entropy losses on sequence i. Define the challenger's paired advantage as

$$d_i = L_i^K - L_i^C. \tag{1}$$

A positive d_i indicates that the challenger assigns higher likelihood to the observed tokens. The sample mean advantage is

$$\widehat{\mu} = \frac{1}{N} \sum_{i=1}^{N} d_i. \tag{2}$$

Pairing removes variation caused by evaluating the two models on different examples and makes the statistical test directly sensitive to their loss difference.

3.2 Bootstrap lower confidence bound

Mean improvement alone is insufficient for promotion. The validator applies a percentile bootstrap [3] to the paired advantages, drawing $B = 10{,}000$ bootstrap samples of the sequence indices with replacement. For replicate b, it computes

$$\widehat{\mu}^{*(b)} = \frac{1}{N} \sum_{j=1}^{N} d_{i_j^{(b)}}, \qquad b = 1, \ldots, B. \tag{3}$$

The lower confidence bound is the empirical α-quantile of the bootstrap distribution,

$$\mathrm{LCB} = Q_\alpha(\widehat{\mu}^*), \qquad \alpha = 0.001. \tag{4}$$

The challenger is crowned if

$$\mathrm{LCB} > \delta, \qquad \delta = 0.0015, \tag{5}$$

using the default values reported for the competition. Thus promotion requires both evidence of positive mean improvement and a minimum effect size. The strict quantile reduces sensitivity to noisy, marginal wins, although repeated adaptive submissions still warrant separate multiple-testing analysis. Algorithm 1 summarizes the procedure, and Table 1 consolidates the parameters that govern it.

3.3 Evaluation-sample selection

Evaluation sequences are selected algorithmically from the submission block hash and the active data-source mixture. The submitting miner does not select the examples, and both models receive identical sequences. This design limits opportunities for favorable sample selection. It does not by itself rule out training–evaluation overlap, benchmark contamination, or strategic adaptation to a publicly known data distribution; these require independent audits.

Algorithm 1 Evaluation of a single challenge (duel)

Require: king parameters $\theta^{(K)}$, challenger parameters $\theta^{(C)}$, submission block hash h, active data-source mixture w, sequence count N, bootstrap replicates B, significance level α, threshold δ
Ensure: decision: crown the challenger, or retain the king

$S \leftarrow \textsc{DeterministicSample}(h, w, N)$ ▷ N sequences of 2,048 tokens, identical for both models
for $i = 1$ to N **do**
  $L_i^K \leftarrow$ mean cross-entropy of $\theta^{(K)}$ on S_i
  $L_i^C \leftarrow$ mean cross-entropy of $\theta^{(C)}$ on S_i
  $d_i \leftarrow L_i^K - L_i^C$
end for
$\widehat{\mu} \leftarrow \frac{1}{N}\sum_{i=1}^{N} d_i$
for $b = 1$ to B **do**
  draw indices $i_1, \ldots, i_N$ uniformly with replacement from $\{1, \ldots, N\}$
  $\widehat{\mu}^{*(b)} \leftarrow \frac{1}{N}\sum_{j=1}^{N} d_{i_j}$
end for
$\mathrm{LCB} \leftarrow Q_\alpha\big(\{\widehat{\mu}^{*(b)}\}_{b=1}^{B}\big)$
if $\mathrm{LCB} > \delta$ **then**
  return crown the challenger
else
  return retain the king
end if

3.4 Selection in practice

Figure 2 shows every completed duel retained in the public validator record at the time of writing: 2,218 records spanning June 2 to August 14, of which 2,117 fall on or before the August 10 report cutoff (98% of the 2,163 completed duels; the rolling history has since discarded a small number of the earliest events). Three properties of the mechanism are visible directly in the data. First, selection is strict: only 196 of the 2,117 retained duels (9.3%) cleared the acceptance boundary, and the cloud of rejected challengers sits overwhelmingly below the threshold line rather than just under it, indicating that most rejected challengers were not near-misses. Second, the acceptance margin is a live control surface: δ was 0.0025 for the first seven weeks and was lowered to 0.0015 on July 21, and the evaluation-sample count was raised twice, reaching $N = 25{,}000$ from July 24. Third, the incumbent's evaluation loss (Figure 2b) falls as a staircase of small accepted improvements punctuated by occasional larger drops, with visible level shifts when the evaluation mixture itself was reweighted.

In addition to the completed duels, roughly one thousand further submissions in the retained window failed pre-duel validation (missing or unreachable digests, malformed configurations, ownership-token mismatches) and were rejected without consuming evaluation compute.

4 Incentives and Global Control

4.1 Five-checkpoint reward window

Subnet incentives are allocated to the current king and up to four preceding kings whose hotkeys remain registered. When all five are present, each receives 20% of the allocated weight. If a historical king leaves the metagraph, shares are renormalized among the eligible checkpoints. A successful miner therefore continues receiving rewards for as many as four subsequent dethronements.

This mechanism softens a purely winner-takes-all contest while preserving a high entry threshold: a model enters the reward window only by becoming king. It also values checkpoints as steps in a lineage, rather than treating them as worthless immediately after a small subsequent improvement. The design may, however, favor frequent incremental coronations or strategic identities; the extent of such behavior is not quantified in the available competition record.

4.2 Global hyperparameters

Although miners control local training, validators can steer the global objective through the acceptance margin δ, evaluation-sample count, and dataset weights. Sixteen configuration updates were made during the competition: 15 changes to dataset configuration, one change to δ, and two changes to the evaluation-sample count (some updates changed more than one quantity).

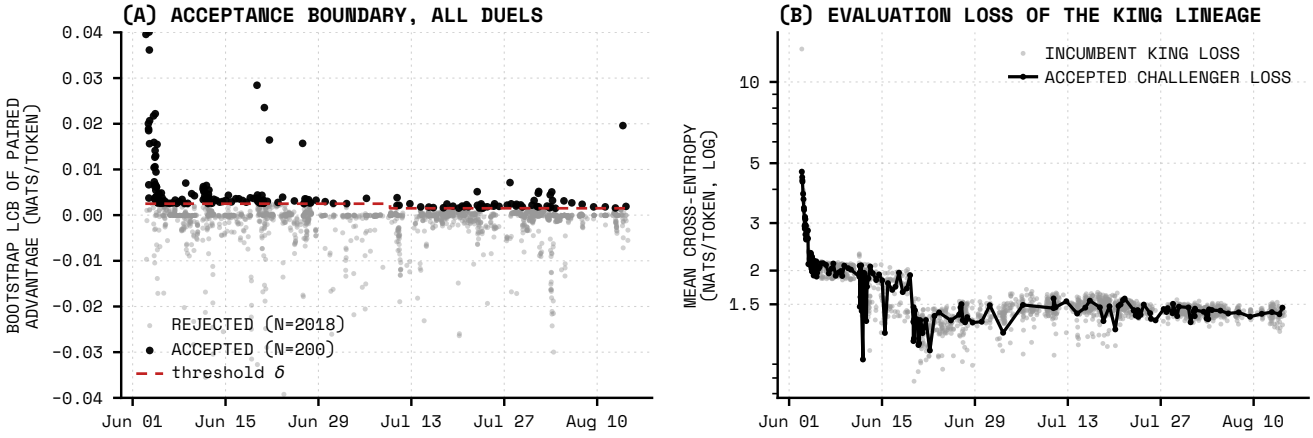

Figure 2: The selection mechanism, from the public validator record (2,218 completed duels, June 2–August 14, 2026). **(a)** Bootstrap lower confidence bound of the paired advantage for every completed duel. The dashed step line is the acceptance threshold δ, lowered from 0.0025 to 0.0015 on July 21. Points are clipped to ± 0.04 nats/token for legibility. **(b)** Mean evaluation cross-entropy of the incumbent king at each duel (grey) and of each accepted challenger (black staircase). Level shifts reflect changes to the evaluation data mixture rather than model regressions.

This creates a two-level optimization process: miners search locally for better checkpoints, while the subnet adjusts which improvements count as globally valuable. Both levers are directly visible in the duel record (Figure 2).

5 Experimental Record

5.1 Competition scale and trajectory

Table 2 summarizes the run. The initial randomly initialized model averaged 30.56% on the reported benchmark aggregate. Within seven days, the score reached 56.48%, a gain of 25.92 percentage points and more than 80% of the total improvement to the later peak. Training emphasis shifted toward mathematics and reasoning in the second half of June; performance subsequently plateaued near 61–62% during July. King #191 reached the peak of 62.28% on August 1. The final configuration point was recorded on August 10.

Figure 3 plots the full benchmark trajectory of the king lineage: the unweighted 11-task mean of every king for which the daily benchmark service (built on the lm-evaluation-harness [4]) completed a full evaluation (94 checkpoints between June 6 and August 14). The earliest benchmarked king, four days into the competition, already averaged 55.8%, confirming that most of the headline gain was realized in the opening week; the remaining ten weeks added roughly six points through many small accepted improvements. The lineage crossed the reported aggregate of Covenant 72B in mid-June and that of Quasar-Preview 18B by early July.

5.2 Participation

The retained duel record contains submissions from 73 distinct coldkeys publishing under 22 distinct repository namespaces. Figure 4 shows the flow of duels and coronations over time and the concentration of coronation events across miner identities. Participation is broad at the duel level but concentrated at the coronation level: the most successful namespace accounts for roughly 39% of recorded coronation events, and the top three for about 63%. This is consistent with the incentive analysis above: the reward window pays only kings, so persistent, well-resourced miners dominate the accepted lineage while a long tail of participants probes the boundary.

5.3 Data mixtures

Across the run, 12 named datasets were active for different durations (Table 3). The evaluation corpus was reported to contain more than four trillion tokens, with samples drawn from the active mixture. Dataset introduction, removal, and reweighting were used to redirect training toward desired capabilities. Active-day counts describe inclusion in the global objective, not the number of tokens consumed by individual miners.

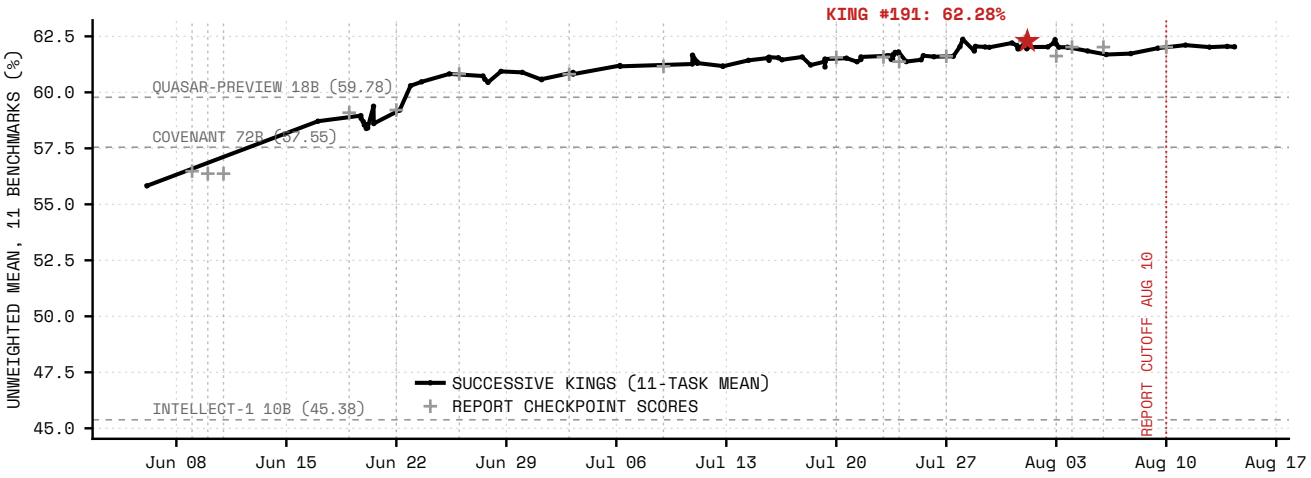

Figure 3: Benchmark trajectory of the king lineage. Each solid marker is one crowned checkpoint with a completed 12-benchmark evaluation from the daily benchmark service (94 of the crowned kings; the seed model, at 30.56%, is off-scale). Dotted vertical lines mark the 16 global hyperparameter updates, and grey crosses are the checkpoint scores published in the technical report's appendix; the two independently produced series agree closely. Dashed horizontal lines are the reported 11-task means of comparison models from Table 6. The red star marks king #191, the checkpoint released as Teutonic-I 10B.

Table 4 reports the final evaluation mixture as published in the validator's dataset manifest (August 13, 2026): 11 enabled sources totalling 4.24 trillion pre-tokenized evaluation tokens, with per-source sampling weights. General web text (FineWeb-Edu [10]) retains the largest single weight, but mathematics and reasoning sources jointly account for roughly half of the sampling mass, reflecting the mid-competition steer toward those capabilities.

6 Downstream Evaluation

6.1 Evaluation protocol

Table 5 states the protocol under which the Teutonic-I 10B scores were produced: nine of the 11 shared benchmarks are evaluated zero-shot, BBH uses three in-context examples, and MATH-500 uses four. `acc` is accuracy from the log-likelihood assigned to each answer option; `acc_norm` normalizes that log-likelihood by option length so longer options are not systematically penalized; `exact_match` is string-level agreement of the extracted final answer. For comparison models, the scores in Table 6 are the best figures published by their respective authors, or measurements under this protocol where none were published. Externally reported scores were not necessarily produced under this protocol; in particular, the settings behind Quasar-Preview 18B's MATH-500 score are not documented.

6.2 Shared-benchmark comparison

Table 6 compares reported scores on 11 benchmarks available for all six models: MMLU [5], ARC-C and ARC-E [2], PIQA [1], HellaSwag [20], OpenBookQA [8], BBH [15], TruthfulQA [7], WinoGrande [14], GPQA [13], and MATH-500 [6]. The comparison set spans other decentralized training initiatives, including INTELLECT-1 10B [11] and Psyche Consilience 40B [9]. The final row is an unweighted arithmetic mean; it does not account for benchmark difficulty, test-set size, uncertainty, or potential differences in evaluation harnesses. On this aggregate, Teutonic-I 10B scores 62.28%, ahead of Quasar-Preview 18B at 59.78% and Covenant 72B at 57.55%. It leads 8 of 11 rows. Quasar-Preview leads PIQA and MATH-500, while Covenant leads HellaSwag.

The largest notable advantage over Covenant is on MMLU, where Teutonic-I 10B scores 75.29% versus 67.11%. Teutonic-I also leads the comparison on both ARC tasks, BBH, TruthfulQA, WinoGrande, GPQA, and OpenBookQA. Its clearest relative weakness is MATH-500: 39.40% compared with 71.40% for Quasar-Preview. This heterogeneity is obscured by the mean and argues against interpreting the aggregate as a complete model ranking.

On the separately reported MMLU-Pro [19] evaluation (five in-context examples, scored with `exact_match` after a dedicated answer-extraction filter), Teutonic-I 10B achieved 39.84%, versus 33.20% for Quasar-Preview 18B. Because comparable MMLU-Pro scores were not provided for all models, this result is excluded from the 11-task mean.

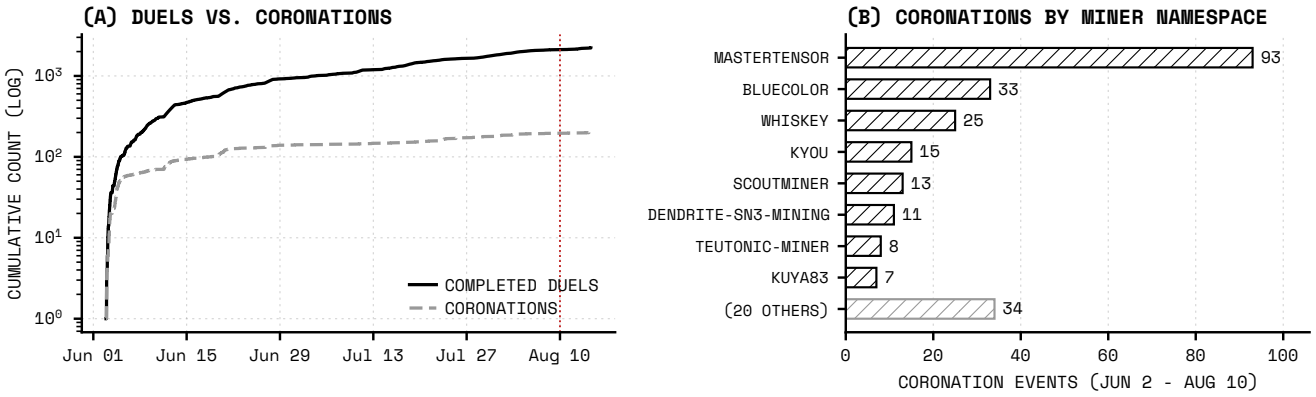

Figure 4: Participation in the competition. **(a)** Cumulative completed duels and coronations (log scale) in the retained public record; the dotted line marks the August 10 report cutoff. **(b)** Coronation events by miner repository namespace. Event counts derive from the public benchmark index, which also records re-coronations after validator state restores, so they slightly exceed the 203 deduplicated official reigns.

Table 2: Summary of the reported SN3 pretraining competition.

Quantity	Value
Duration	70 days
Competition dates	Jun. 2–Aug. 10, 2026
Architecture	Quasar-10B
Parameter count	10B
Context length	2,048 tokens
Completed duels	2,163
Successful coronations	203
Datasets used	12
Hyperparameter updates	16
Acceptance rate	9.3%
Median duel wall time	17.7 min
Distinct submitting coldkeys	73
Best checkpoint	King #191
Date of best checkpoint	Aug. 1, 2026
Best 11-task mean	62.28%

6.3 Per-benchmark dynamics

Figure 5 decomposes the aggregate trajectory into its 12 constituent benchmarks. The dynamics are heterogeneous. Knowledge- and reading-oriented tasks (MMLU, ARC-E, PIQA, HellaSwag, WinoGrande) rose quickly and then drifted upward slowly, closely tracking the pretraining-loss staircase of Figure 2b. MATH-500 instead moved in discrete jumps aligned with the introduction and reweighting of mathematics corpora in the global mixture, from near zero to just under 40%. TruthfulQA and GPQA improved modestly and noisily, as expected for tasks weakly coupled to next-token loss on the evaluation mixture. This decomposition illustrates both the strength and the limit of loss-based selection: capabilities well represented in the evaluation mixture improve steadily, while others move only when the mixture is steered toward them.

6.4 Interpretation

The comparison demonstrates that a 10B checkpoint produced by the SN3 process can exceed several larger models on this particular suite and score aggregation. It does not demonstrate that the model is uniformly superior, more compute efficient, or more capable in deployment. Parameter count is only one factor; training tokens, data quality, contamination, optimizer state, total network compute, and evaluation implementation may all affect the outcome. A stronger study would reevaluate every checkpoint in one harness, report confidence intervals, include contamination checks, and normalize by total training FLOPs and monetary cost.

Table 3: Reported number of active days for each data source during the 70-day competition. Content spans education, mathematics, synthetic data, mixed content, code, logic, economics, moral scenarios, science, scientific articles, multiple domains, and WikiHow-style material.

Dataset	Days	Dataset	Days
FineWeb-Edu	70	AutoMathText-V2	62
UltraData-Math / L3	62	Quasar-SN3	29
Nemotron Specialized v1.1	36	Nemotron Specialized v1.2	53
Nemotron-CC-Math	46	OpenThoughts3-1.2M	46
peS2o-v3	31	OpenMathReasoning	22
Dendrite-Synth v1/v2/v3	20	Cosmopedia	20

Table 4: Final evaluation mixture from the published dataset manifest (Aug. 13, 2026). Weights are sampling probabilities over sources; token counts are pre-tokenized evaluation tokens available per source.

Source	Weight	Tokens
FineWeb-Edu	0.25	1.58T
AutoMathText-V2	0.19	2.25T
OpenMathReasoning	0.10	39.2B
Dolma3 LongMino (8k pool)	0.10	70.6B
UltraData-Math-L3	0.08	78.0B
OpenThoughts3-1.2M	0.08	18.7B
peS2o-v3	0.06	79.6B
Cosmopedia (WikiHow/stories)	0.06	2.8B
Nemotron-CC-Math v1 (4+, MIND)	0.04	76.0B
Nemotron Specialized v1.2	0.03	43.6B
Dendrite-Synth	0.01	0.6B
Total	1.00	4.24T

7 Discussion

Decentralization as search. SN3 decentralizes experiment selection rather than a single gradient calculation. This may broaden the search over curricula and optimization recipes, allow specialized participants to contribute independently, and make progress measurable through an auditable incumbent. Sequential inheritance also lets later miners build on earlier improvements.

Selection pressure. The combination of paired evaluation, a low bootstrap quantile, and a positive margin discourages noisy replacement. Yet a fixed objective can invite over-specialization to its data mixture. Changes to global dataset weights can counter stagnation, but they also make the target nonstationary and place substantial influence in the hands of whoever controls those weights.

Compute efficiency. The system may avoid committing all resources to one centrally chosen recipe, but rejected challengers consume compute without directly changing the model. Without miner-level token and hardware logs, total compute efficiency cannot be compared with centralized pretraining. The relevant question is not merely whether the final model is small, but how much aggregate computation and data were required to discover it. The validator-side cost is measurable: at a median of 17.7 minutes per duel on one eight-GPU server, the full 2,163-duel record represents roughly 640 GPU-hours of evaluation compute, small relative to any plausible estimate of the miners' aggregate training compute.

Reproducibility. Deterministic sample selection and explicit promotion rules make individual duels more reproducible than subjective model claims. End-to-end reproduction still requires versioned checkpoints, exact data-mixture snapshots, evaluation code, random seeds, tokenizer details, miner training disclosures, and a record of every global configuration change.

8 Limitations and Broader Impacts

This report is based on the competition statistics and benchmark values made available by the project. We did not independently rerun the benchmarks or verify every checkpoint lineage. The comparisons may mix evaluation harnesses or prompting

Table 5: Evaluation protocol for the 11 shared benchmarks.

Benchmark	Shots	Metric	Benchmark	Shots	Metric
MMLU	0	`acc`	OpenBookQA	0	`acc_norm`
ARC-C	0	`acc_norm`	BBH	3	`acc_norm`
ARC-E	0	`acc_norm`	TruthfulQA	0	`acc`
PIQA	0	`acc_norm`	WinoGrande	0	`acc`
HellaSwag	0	`acc_norm`	GPQA	0	`acc_norm`
MATH-500	4	`exact_match`			

Table 6: Reported benchmark scores (%). Higher is better. Bold denotes the best value in each row. "Psyche" abbreviates Psyche Consilience.

Benchmark	Teutonic-I 10B	Quasar 10B	Quasar-Preview 18B	INTELLECT-1 10B	Psyche 40B	Covenant 72B
MMLU	**75.29**	49.63	60.87	32.69	24.23	67.11
ARC-C	**63.82**	41.89	63.40	44.80	31.14	56.83
ARC-E	**84.97**	62.96	82.45	71.76	55.77	80.93
PIQA	82.81	69.91	**83.30**	77.73	76.12	81.56
HellaSwag	79.42	62.37	73.07	70.26	63.67	**80.61**
OpenBookQA	**49.00**	35.40	46.40	43.80	35.20	44.00
BBH	**49.51**	31.26	38.10	32.93	30.50	45.96
TruthfulQA	**49.58**	41.01	41.70	35.45	37.90	49.41
WinoGrande	**77.35**	56.27	67.56	63.30	56.99	75.85
GPQA	**33.98**	25.84	29.28	25.84	24.66	30.03
MATH-500	39.40	0.00	**71.40**	1.00	0.20	20.80
Unweighted mean	**62.28**	43.32	59.78	45.38	39.67	57.55

conventions, and no uncertainty estimates are available for the downstream scores. The unweighted mean treats all benchmarks equally and can be strongly affected by outliers such as MATH-500. There is no matched centralized baseline trained with the same architecture, data access, token budget, and aggregate compute, so causal claims about the benefit of decentralization are not supported. The figures in this paper are computed from the public validator record, which is a rolling window: a small fraction of the earliest duel events had already been discarded at the time of analysis, and the benchmark index records coronation events rather than deduplicated reigns.

Open competition can broaden participation in model development and create transparent, performance-based incentives. Conversely, financial rewards may encourage benchmark gaming, checkpoint copying, hidden data use, or duplicated compute. Data provenance, licensing, privacy, model safety, and the energy cost of unsuccessful challengers require governance beyond loss-based selection. Future competitions should include provenance attestations, contamination audits, compute reporting, safety evaluations, and mechanisms for detecting copied or minimally modified submissions.

The subnet operators state that the next competition will target substantially larger models, potentially around 100B parameters, and that mechanisms developed during this run — synthetic-dataset generation and autonomous selection of pretraining hyperparameters among them — will be documented separately.

9 Conclusion

The Teutonic Subnet 3 competition produced a strong 10B language model through 70 days of sequential, decentralized checkpoint improvement. A statistically conservative duel mechanism converted independent training experiments into a shared model lineage, yielding 203 accepted improvements from 2,163 completed challenges. The best checkpoint averaged 62.28% over 11 reported shared benchmarks and ranked first on 8, exceeding the reported aggregate of several models with more parameters. The result is evidence that competitive checkpoint selection is a viable coordination mechanism for pretraining research. Establishing its efficiency and general advantage over centralized training will require controlled baselines, standardized reevaluation, aggregate-compute accounting, and stronger data and checkpoint audits.

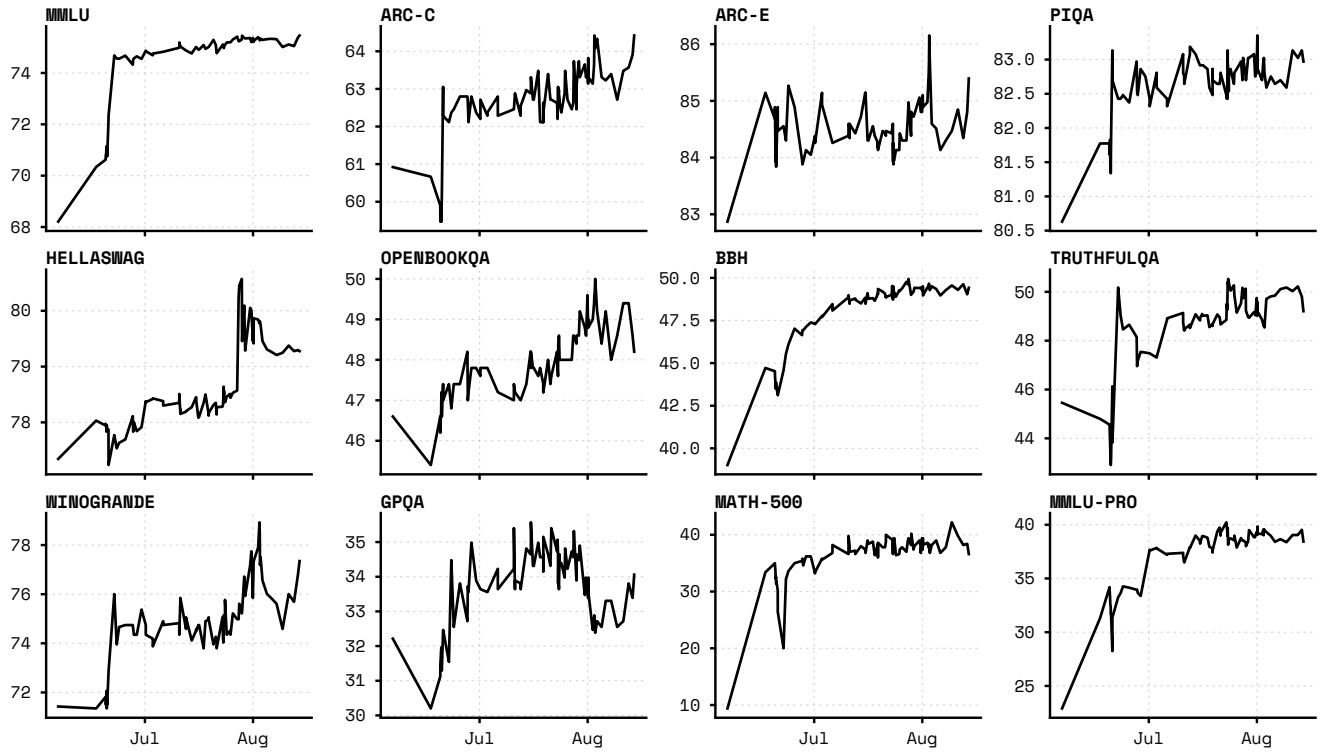

Figure 5: Per-benchmark scores of successive kings (June 6–August 14, 2026), from the daily benchmark service. Note the differing vertical dynamics: web-knowledge tasks improve smoothly, while MATH-500 moves in steps aligned with changes to the evaluation data mixture.

Declaration of Generative AI Use

The design and operation of the subnet, the pretraining competition, and all measurements and benchmark results are the work of the subnet's operators and participants. A generative AI assistant was used in the preparation of this manuscript: to structure and rewrite the prose in an academic register, to extract and analyze the public validator record, to generate the figures from that data, and to typeset the document in LaTeX.

Reproducibility Statement

The system implementation is available through the Teutonic repository, the released model weights are hosted on Hugging Face (`https://huggingface.co/dendriteholdings/Teutonic-I`), and project information and further artifacts are linked below. A complete release should preserve all evaluated checkpoints, block-derived sample identifiers, data-mixture configurations, validator code, and benchmark-harness versions. The headline numerical values in this paper are transcribed from the supplied SN3 technical report. Figures 2–5 and Tables 2 (acceptance rate, wall time, participation rows) and 4 are computed directly from the public validator artifacts as retrieved on August 14, 2026: the live dashboard state (`dashboard.json`), the all-kings benchmark index (`king-benchmark-daily/all-kings/index.json`), and the dataset manifest (`dataset/all-datasets.manifest.json`), all served from the project's public object store. The extraction and plotting script is included alongside this manuscript.

References

[1] Y. Bisk, R. Zellers, R. Le Bras, J. Gao, and Y. Choi. PIQA: Reasoning about physical commonsense in natural language. In *AAAI*, 2020.

[2] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. Think you have solved question answering? Try ARC, the AI2 Reasoning Challenge. *arXiv:1803.05457*, 2018.

[3] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[4] L. Gao, J. Tow, B. Abbasi, et al. A framework for few-shot language model evaluation. Zenodo, https://doi.org/10.5281/zenodo.5371628, 2023.

[5] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021.

[6] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks*, 2021. MATH-500 denotes the 500-problem subset introduced by Lightman et al., *Let's Verify Step by Step*, 2023.

[7] S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *ACL*, 2022.

[8] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *EMNLP*, 2018.

[9] Nous Research. *Psyche: a decentralized training network*. https://nousresearch.com/nous-psyche/, 2025.

[10] G. Penedo, H. Kydlíček, L. Ben allal, A. Lozhkov, M. Mitchell, C. Raffel, L. Von Werra, and T. Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *NeurIPS Datasets and Benchmarks*, 2024.

[11] Prime Intellect. INTELLECT-1 technical report. *arXiv:2412.01152*, 2024.

[12] Y. Rao. *Bittensor: A peer-to-peer intelligence market*. Bittensor whitepaper, https://www.bittensor.com/whitepaper.

[13] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. *arXiv:2311.12022*, 2023.

[14] K. Sakaguchi, R. Le Bras, C. Bhagavatula, and Y. Choi. WinoGrande: An adversarial Winograd schema challenge at scale. In *AAAI*, 2020.

[15] M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, and J. Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of ACL*, 2023.

[16] Teutonic contributors. *Teutonic: decentralized model pretraining infrastructure*. https://github.com/unarbos/teutonic, 2026.

[17] Teutonic. *Teutonic project website*. https://teutonic.ai/, 2026.

[18] Teutonic contributors. *Teutonic-I 10B model weights*. https://huggingface.co/dendriteholdings/Teutonic-I, 2026.

[19] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *NeurIPS Datasets and Benchmarks*, 2024.

[20] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. HellaSwag: Can a machine really finish your sentence? In *ACL*, 2019.

A Configuration-checkpoint scores

Table 7 lists the average benchmark score of the reigning king at each configuration checkpoint, as published in the technical report. Each score was computed for the final king preceding a given hyperparameter update; when the corresponding king checkpoint was unavailable, the closest matching model was used instead. The June 2 entry is the randomly initialized starting model, and the August 1 entry is king #191. These are the green crosses overlaid on Figure 3.

Table 7: Average benchmark score of the reigning king at each configuration checkpoint (day.month, 2026).

Date	Score	Date	Score
02.06	30.56	20.07	61.56
09.06	56.48	23.07	61.58
10.06	56.37	24.07	61.37
11.06	56.37	27.07	61.61
19.06	59.09	01.08	62.28
22.06	59.21	03.08	61.62
26.06	60.82	04.08	62.02
03.07	60.81	06.08	62.02
09.07	61.16	10.08	62.02