Teutonic-I audit

> **TL;DR**
>
> **Teutonic-I was not pretrained from scratch by Bittensor miners and shows no aggregate benchmark gain over its Qwen base.**

Overview

The Teutonic-I report presents the model as the product of a decentralized pretraining competition that began from a random initialization, and compares it with models described as decentralized, from-scratch training efforts. It discloses adopting Quasar-10B's *architecture and tokenizer*, but not that pretrained Quasar or Qwen weights themselves entered the competition. The weights show something categorically stronger: the released checkpoint descends from Quasar's *weights* and, through them, from Qwen3.5-9B-Base, a model pretrained by Alibaba at tens-of-trillions-of-tokens scale.

Quasar-10B's own model card explicitly states that it was initialized from Qwen3.5-9B-Base. The numerical chain is equally clear: Quasar and Qwen correlate at **0.9999**, while Teutonic and Quasar correlate at **0.9596**. Notably, a checkpoint correlating at 0.9996 with Quasar was crowned two days into the competition.

The Teutonic-I report does not compare Teutonic-I with its Qwen source model nor mention the source Qwen model. In the matched audit, Teutonic-I produces no gain in the unweighted 11-task average; the results are consistent with recovery of performance lost after Quasar introduced its replacement attention components.

Teutonic-I is derived from Qwen3.5-9B-Base

We use a simple correlation method to compare the model weights: Pearson correlation. When a weight is relatively large, small, positive, or negative in one model, Pearson correlation checks whether the corresponding weight tends to follow the same pattern in the other model. A value near **1** means extremely similar numerical structure; a value near **0** means no aligned relationship. Merely sharing an architecture does not produce a high value: the same-layout random genesis (the competition's published randomly initialized starting checkpoint) is approximately zero, a shape-matched Qwen-family control is 0.00003, and deliberately pairing the right tensor type with the wrong layer gives approximately 0.016.

Against those controls, **0.9596 across all 8.60 billion corresponding parameters is extraordinarily high**. The result is far beyond plausible coincidence and places shared weight ancestry beyond reasonable doubt.

Comparison	Compared values	Pearson r	Exact BF16 equality
Qwen3.5 ↔ Quasar	8.484B	**0.999903**	78.48%
Quasar ↔ Teutonic	8.602B	**0.959555**	0.7868%
Qwen3.5 ↔ Teutonic	8.484B	**0.960098**	0.7881%
Quasar ↔ layer-shifted Qwen3.5	6.450B	0.015676	—
Pythia-410M: seed1 ↔ seed2	405M	0.000014	0.0892%

The Qwen–Quasar result directly exposes Quasar's disclosed Qwen initialization: the models correlate at 0.999903 and 78.48% of their aligned BF16 values are identical. Teutonic then correlates at approximately 0.96 with both Quasar and Qwen, directly placing it in the same weight lineage. By comparison, deliberately shifting the same tensor type to a different compatible layer collapses the correlation to 0.015676.

The Pythia row shows what independent training actually produces. EleutherAI's PolyPythias provide runs that share the same 410M architecture, tokenizer, training data, and hyperparameters, differing only in random seed — yet two such runs correlate at 0.000014, with BF16 equality at the chance rate. If even identical data and configuration yield zero alignment, two models trained on different data cannot reach 0.96 through training alone; correlation at that level implies descent from shared weights.

AWM cross-check

We also applied AWM, a fingerprinting method designed specifically to detect model lineage. It compares the geometry of attention query and key weights and is less dependent than Pearson on exact coordinate-by-coordinate equality. Values near 1 indicate strongly aligned weight structure consistent with shared lineage; values near 0 indicate no detected alignment.

Comparison	Median layer alignment (mean Q/K UCKA)
Qwen3.5 ↔ Quasar	**0.999923**
Quasar ↔ Teutonic-I	**0.949501**
Qwen3.5 ↔ Teutonic-I	**0.949391**
Random init ↔ Quasar	-0.000043
Pythia-410M: seed1 ↔ seed2	0.002013

The AWM check reaches the same conclusion as Pearson: the disclosed Qwen–Quasar link is almost maximal, both comparisons involving Teutonic remain near 0.95, and the same-layout random model and the independently trained Pythia seed pair are both effectively zero. This confirms that Teutonic-I is a close derivative of Qwen3.5-9B-Base through Quasar-10B.

What is Qwen3.5-9B-Base?

Qwen3.5-9B-Base is a 9B-parameter base model developed by Alibaba and the Qwen team through conventional, centrally coordinated data-center pretraining. Qwen has not disclosed an exact training-token count for this checkpoint. Its family-level materials say Qwen3.5 training used a larger scale of visual-text tokens than Qwen3, which used 36T tokens. A reasonable conservative estimate for Qwen3.5-9B is 36T tokens; against Covenant-72B's 1T figure, that gives **36× as much training data.**. Even though Qwen3.5-9B is eight times smaller, it uses at least **5× the training compute**.

What is Quasar-10B?

Quasar-10B is a conversion of Qwen3.5-9B-Base posted on huggingface by SILX. It retains most of Qwen's weights but replaces eight attention layers with gated linear attention (GLA) modules. Those replacement modules contain **118,000,128** of Quasar's **8,602,037,248** parameters, or approximately **1.37%** of the model. Their learned gates regulate the internal update, while each module enters the model through a residual branch whose overall contribution can also be scaled. Poor calibration can therefore let these new branches distort the inherited Qwen computation; reducing their influence can partially recover it. The public Quasar-10B checkpoint heavily degrades performance on top of the base Qwen model.

Notably a recent weight-lineage audit reports a similar pattern for Quasar-Preview 18B also released by SILX: 98.2% of its parameters are bit-identical to Ant Group's Ling-mini-base-2.0-20T checkpoint, while the remaining 1.8% consists of added hybrid-attention parameters. Quasar-Preview's model card does not disclose that Ling checkpoint lineage.

An almost-Quasar checkpoint was crowned near the start

Although the initial model in Teutonic competition was randomly initialized, a checkpoint derived from Quasar-10B and Qwen appears to have entered the competition by at least **June 4, 2026**. That checkpoint, teutonic-5g1dh3iz-test49, was crowned that day. A full comparison found **56.23%** exact BF16 equality and a Pearson correlation of **0.999621** with Quasar-10B.

What the chronology means. The competition accepted architecture-compatible pretrained submissions and did not authenticate or prohibit their use. The June 4 crown shows that this was not merely a theoretical loophole: an almost-Quasar checkpoint won very early in the competition.

The weight evidence helps explain the later results. Quasar left most comparable Qwen weights unchanged — 78.48% are bit-identical and the full aligned correlation is 0.999903 — while adding eight replacement attention modules that appear poorly integrated. The competition therefore appears mainly to have repaired a weak Qwen-derived intermediate rather than produced a new pretrained base. The evidence does not imply that miners performed no subsequent training at all; it shows that the pretraining base was inherited, thus it is not a pre-training competition.

Teutonic shows no aggregate benchmark gain over Qwen

Given the above observation the source Qwen model is a critical comparison missing from the Teutonic report. Here we fill in this missing gap by evaluating Qwen3.5-9B. The results show that Teutonic does not appear to provide a meaningful benefit over its source model.

The two blue columns — Qwen3.5-9B-Base and Teutonic-I — were measured under the same pinned lm-evaluation-harness configuration and tasks as the report. For the benchmark-score rows, the remaining columns reproduce the paper's reported comparisons and do not represent matched reruns. The final perplexity row is a separate fixed-text WikiText-2 audit; lower perplexity is better.

Benchmark	Qwen3.5-9B (audit run)	Teutonic-I (audit run)	Δ T-Q	Teutonic-I (reported)	Quasar 10B	Quasar-Prev. 18B
MMLU 0-shot acc	77.25	75.40	-1.85	75.29	49.63	60.87
ARC-C 0-shot acc_norm	56.66	64.51	+7.85	63.82	41.89	63.40
ARC-E 0-shot acc_norm	78.07	85.02	+6.95	84.97	62.96	82.45
PIQA 0-shot acc_norm	80.52	82.43	+1.91	82.81	69.91	83.30
HellaSwag 0-shot acc_norm	79.25	79.53	+0.28	79.42	62.37	73.07
OpenBookQA 0-shot acc_norm	44.60	49.20	+4.60	49.00	35.40	46.40
BBH 3-shot acc_norm	62.52	49.38	-13.14	49.51	31.26	38.10
TruthfulQA 0-shot acc (mc2)	51.79	49.63	-2.16	49.58	41.01	41.70
WinoGrande 0-shot acc	73.32	77.03	+3.71	77.35	56.27	67.56
GPQA 0-shot acc_norm (main)	41.07	36.38	-4.69	33.98	25.84	29.28
MATH-500 4-shot exact_match	42.20	38.40	-3.80	39.40	0.00	71.40
Average score (11 tasks)	**62.48**	**62.45**	**-0.03**	**62.28**	**43.32**	**59.78**
WikiText-2 perplexity fixed-text audit; lower is better	11.5	11.5	0.0	—	20.4	—

The matched benchmark result is straightforward: Qwen scores **62.48** and Teutonic **62.45** on the average across 11 tasks, a difference of **-0.03**. Teutonic leads six rows and Qwen leads five, sometimes by material margins, so the near-equal average reflects task rebalancing rather than uniform equivalence. It nevertheless provides no measured aggregate improvement over Qwen. On a WikiText-2 sample, Teutonic and Qwen both achieve **11.5** perplexity, while Quasar is substantially worse at **20.4**. Thus Teutonic-I does not show significant improvement.

The paper-reported columns place Quasar-10B at **43.32** and Teutonic-I at **62.28**. This illustrates how weak the Quasar checkpoint was in the paper's own comparison.

Why Quasar may have underperformed

Quasar inherited most of its weights from Qwen but replaced eight attention layers. The inherited projections in those replacement blocks remain extremely close to Qwen, while the new gate-related tensors retain initialization-like statistics and every gated-normalization scale remains at its initial value of 1.0. This is consistent with a poorly calibrated conversion and offers a plausible explanation for Quasar's weak reported results. Because the original Qwen weights remain embedded in the checkpoint, recovering Qwen-level behavior is much easier than learning comparable behavior from an early-training model. In a perplexity evaluation, downweighting only the Quasar-specific branches improved Quasar's likelihood, although that intervention alone did not fully close the gap to Qwen.

What can reasonably be concluded

- **Teutonic-I was not pretrained from scratch by Bittensor miners.** Its released weights clearly descend from Qwen3.5-9B-Base through the Quasar weight family. Thus, it is not comparable to Intellect-1, Psyche-40B, and Covenant-72B.
- **The Teutonic competition produced no clear aggregate improvement over the heavily pretrained Qwen base used by miners.** Final Teutonic mainly returned to Qwen-level performance after the much weaker Quasar conversion.

Sources

- Teutonic-I technical report and public code
- Quasar-10B model card
- Qwen3.5-9B-Base model card
- AWM model-provenance paper and official implementation