NexisGen: Decentralized Training Dataset Production on Bittensor

Subnet 70 | Bittensor | Version 1.0 - April 2026

Abstract

Every frontier AI model is only as good as the data it trains on. OpenAI, Google, and Anthropic spend hundreds of millions building proprietary data pipelines - and keep the results locked up. The open-source alternatives are cheap but unreliable: inconsistent quality, no provenance, no accountability.

NexisGen is a decentralized training data network on Bittensor. Independent miners compete to produce high-quality, annotated datasets - video, audio, image, and text - while validators cryptographically verify every sample before it counts. Quality is not a guideline. It is enforced at the protocol level. Miners who produce verified, original data earn TAO emissions. Miners who produce anything less earn nothing.

The protocol is modality-agnostic. It launches with video and scales to every data type that frontier models need. The output is a growing, multi-modal training corpus with provenance guarantees that no centralized provider and no open dataset project can match. The data economy should not be locked inside three companies. NexisGen opens it up - and pays the people who do the work.

1. Problem

1.1 The Data Bottleneck

Frontier AI models are data-hungry. Video understanding, image generation, speech synthesis, multimodal reasoning - every capability traces back to the same dependency: massive volumes of high-quality, annotated training data. The supply side has not kept up.

Manual curation is expensive. Automated pipelines produce noise. The few large-scale datasets that exist are either proprietary - locked inside labs at Google, OpenAI, and Meta - or academically licensed with restrictions that limit commercial use. Video datasets are the most acute bottleneck today, but audio, image, and text datasets face the same problems: inconsistent quality, unclear provenance, no economic incentive to improve. Labs that cannot source good data fall behind. Labs that can source it keep it to themselves.

1.2 Quality Without Accountability

Open datasets - video, audio, image, text - all share the same structural failure: nobody gets paid for quality, and nobody gets penalized for junk. A contributor can upload poorly captioned, low-resolution,

or duplicate content with zero consequence. Without economic skin in the game, quality degrades as the easiest content floods the collection.

The pattern plays out the same way every time. A dataset launches with high standards. Contributors trickle in. Quality control is manual and cannot scale. Within a year, the signal-to-noise ratio has collapsed and the maintainers are overwhelmed. LAION, Common Voice, LibriSpeech derivatives - the story repeats across modalities.

1.3 Provenance and Integrity

Training data provenance matters for compliance, reproducibility, and trust. Most datasets cannot answer basic questions: Where did this clip come from? Has it been tampered with? Is the caption accurate? Are there duplicates? These are not academic concerns. They determine whether a dataset is usable for production model training.

1.4 Design Principles

- **Quality through incentives.** Miners earn proportionally to verified quality. Cubic scoring amplifies the gap between good and mediocre work.
- **Verification over trust.** Every claim a miner makes - hash integrity, resolution, caption accuracy, content category - is independently verified.
- **Deterministic fairness.** Sampling is seeded and reproducible. No validator can selectively target or favour specific miners.
- **Deduplication at protocol level.** Cross-miner overlap detection and global indexing prevent the network from paying twice for the same data.

2. Solution

2.1 A Decentralized Dataset Factory

NexisGen converts decentralized compute and curation effort into verified, training-ready datasets. The protocol defines dataset specifications - pluggable modules that specify schema, asset requirements, and validation rules for a given data type. At launch, the video_v1 spec targets captioned video clips: miners collect video from public sources, segment it into standardized 5-second clips at 1280x720, generate captions using vision-language models, and upload structured dataset packages. Future specs cover audio transcription, image captioning, text curation, and cross-modal alignment - all flowing through the same production and validation framework.

Model trainers get data they can trust. Miners get paid for producing it.

2.2 Interval-Based Production

The network runs on a fixed cadence synchronized to the Bittensor blockchain. Every 100 blocks, miners submit a complete dataset package. Validators evaluate after the interval closes. Weights update every 300 blocks. This is not a one-time data dump. It is a factory that runs around the clock.

2.3 Multi-Layer Validation

Validation is an eight-phase pipeline. Manifest integrity, SHA256 hash verification, resolution enforcement, caption quality, semantic caption-to-frame alignment, category classification, and cross-network deduplication all run in sequence. A submission must pass every layer. One failure rejects the entire interval.

Partial credit encourages partial effort. NexisGen does not do partial credit.

2.4 Cubic Reward Scaling

Miners are scored using a cubic function of verified sample count. A miner with 3 passing samples scores 27x more than a miner with 1. The economics push hard toward maximizing the number of quality clips per interval rather than submitting the minimum viable package.

2.5 Trustless Discovery

Miners publish storage credentials on-chain via Bittensor's commitment system. Validators discover miners and access their data without off-chain coordination, trusted registries, or centralized APIs. Discovery and verification run entirely on on-chain commitments and direct storage access.

3. Architecture

3.1 Participants

Role	Responsibility
Miner	Collects source material for active specs, generates standardized samples and annotations, uploads dataset packages to cloud storage, publishes access credentials on-chain. At launch: video clips. Future: audio, image, text.
Validator	Downloads miner submissions, runs the spec-appropriate validation pipeline, computes scores, submits weights to chain.
Owner-Validator	A designated validator that also maintains the global overlap index and curates the verified dataset collection.
Consumer	Downstream users - model trainers, researchers - who access the verified corpus.

3.2 Three-Layer Architecture

Chain Layer (Bittensor). The metagraph handles miner discovery. Commitment storage holds R2 credentials (128-byte on-chain payloads). Weight submission drives TAO distribution. No custom smart contracts. Everything uses Bittensor primitives.

Storage Layer (Cloudflare R2). Each miner operates a dedicated S3-compatible bucket keyed by hotkey. Miners write dataset packages. Validators read and verify using credentials committed on-chain. Two shared buckets - nexis-record-info for the global deduplication index and nexis-dataset for the curated corpus - are maintained by the owner-validator.

Validation Layer. The eight-phase pipeline runs on each validator independently. Deterministic sampling ensures all validators check the same miners and rows for a given interval. No validator-to-validator coordination. Consensus emerges from independent execution of identical logic.

3.3 Data Flow

A single interval:

- Miner detects new interval via block height
- Miner collects videos, segments clips, generates captions via vision-language model
- Miner assembles dataset.parquet + manifest.json + clip and frame assets
- Miner uploads the complete package to its R2 bucket under the interval prefix
- Interval closes (100 blocks elapsed)
- Validator reads metagraph, fetches miner R2 credentials from on-chain commitments
- Validator downloads manifest and dataset from each miner's bucket
- Validator runs the eight-phase validation pipeline
- Validator accumulates scores across intervals
- Every 300 blocks: validator normalizes scores and submits weights to chain
- Bittensor distributes TAO emissions proportional to stake-weighted consensus

3.4 Dataset Package Structure

Every interval submission follows a fixed structure:

```
{interval_id}/
  manifest.json       # Submission metadata, SHA256 of dataset
  dataset.parquet     # All ClipRecord rows with full metadata
  clips/{clip_id}.mp4  # 5-second video clips at 1280x720
  frames/{clip_id}.jpg # First-frame JPEG for each clip
```

The manifest contains protocol version, schema version, spec ID, category, miner hotkey, interval ID, record count, dataset SHA256 hash, and creation timestamp. The Parquet file holds one row per clip with 18 fields covering identity, hashes, dimensions, caption, source provenance, and split assignment.

3.5 Trust Model

At launch, the owner-validator maintains the global overlap index and curates the verified corpus. Section 7 lays out how this decentralizes.

Everything else is trustless. Miner discovery runs through the metagraph. Credential exchange runs through on-chain commitments. Validation runs independently on each validator with deterministic sampling. Weight submission uses Bittensor's native consensus. No trusted intermediary touches the quality loop. When new specs activate - audio, image, text - they inherit this trust model without modification.

4. Evaluation

4.1 Principles

- **Exhaustive verification.** Every claim in a submission - hash integrity, resolution, caption quality, category accuracy - is independently checked. Nothing is taken on faith.
- **Deterministic sampling.** Validators sample miners and rows using seeded hashes. Selection is reproducible, unpredictable to miners, and consistent across validators.
- **Assume adversaries.** Every check is built for miners trying to game the system, not cooperating with it.

4.2 Eight-Phase Validation Pipeline

The evaluation logic is the protocol itself - open-source, deterministic, identical for every miner. Each dataset spec defines its own validation phases. The table below shows the video_v1 pipeline; future specs will define modality-appropriate equivalents.

Phase	Check	Scope
1. Discovery & Fetch	Download manifest and dataset from miner's R2 bucket	All miners
2. Manifest & Identity	Schema validation, hotkey match, interval match, spec compatibility	Per miner
3. Integrity	SHA256 of dataset.parquet vs. manifest hash, record count match	Per miner
4. Hard Checks	YouTube source, clip overlap >= 5s, caption length >= 20 words, no URLs in captions, clip_start <= 5000s	All rows
5. Sampled Asset Verification	Download clips/frames, verify SHA256 hashes, validate 1280x720 resolution via ffprobe	20% of rows, max 3

Phase	Check	Scope
6. Semantic Caption Check	Vision-model verification that caption matches frame content, prompt injection detection	Sampled rows
7. Category Validation	Two-stage classification: caption gate + strict 3-frame vision scoring across 7 categories	Sampled rows
8. Overlap Pruning	Global deduplication index + cross-miner arbitration by earliest manifest timestamp	Accepted rows

Phase 4 is absolute: the first violation rejects the entire interval. Miners who ship one bad row alongside ninety-nine good ones still fail. The incentive is to validate your own output before you upload it.

4.3 Scoring

Miner scores follow a cubic power law:

```
score = passed_sample_count ^ 3
```

Passed Samples	Score
1	1
2	8
3	27

A miner who passes 3 sample checks earns 27x the score of a miner who passes 1. You cannot game this by submitting one perfect clip and filling the rest with junk.

4.4 Anti-Gaming Detection

Attack	Defence
Fabricated metadata	SHA256 hash verification of every sampled asset against manifest declarations
Resolution spoofing	ffprobe validation of actual pixel dimensions (must be exactly 1280x720)
Low-effort captions	Minimum 20 words, no URLs, semantic alignment check via vision-language model
Prompt injection in captions	Explicit keyword detection for adversarial prompt patterns
Content recycling	Global overlap index with cross-miner arbitration by creation timestamp
Category manipulation	Two-stage vision-model classification: caption gate + strict 3-frame scoring with nature >= 0.72 and margin >= 0.12

Deterministic sampling means miners cannot predict which rows will be checked. Get all of them right or risk losing the interval.

4.5 Failure Gating

The protocol tracks failures with a lookback window of 1 interval. If a miner failed validation in the immediately preceding interval, their weight drops to zero regardless of history. One bad interval wipes your earnings until you fix it.

The short window is deliberate. Miners who diagnose and fix their issues quickly recover quickly. The system rewards consistent production, not past reputation.

5. Token Economics

5.1 Value Flow

NexisGen produces a tangible asset: verified, captioned training datasets. TAO emissions flow to miners who produce this asset. Miners invest compute, bandwidth, and API costs to produce data. Validators invest compute to verify it. The chain distributes TAO proportional to verified output.

5.2 Costs and Emissions

Flow	Mechanism
TAO emissions (from Bittensor)	Distributed to miners proportional to stake-weighted validator consensus on quality scores
Miner costs	R2 storage, OpenAI/Gemini API for captioning, compute for video processing, bandwidth
Validator costs	R2 reads, OpenAI/Gemini API for semantic/category verification, compute for validation pipeline

Miners who produce more verified clips per interval earn more TAO. Miners who produce zero verified clips earn zero TAO. No minimum stake requirement - if you can produce quality data, you can compete.

5.3 Weight Computation

Every 300 blocks, validators compute and submit weights:

```
raw_weight[hotkey] = sum(interval_scores) where no failure in lookback window
normalized_weight[hotkey] = raw_weight[hotkey] / sum(all raw_weights)
```

Weights are submitted as a dense UID-aligned vector via Bittensor's set_weights. The chain aggregates using stake-weighted consensus. Three retry attempts with 10-second backoff handle

transient chain issues.

5.4 Incentive Alignment

Behaviour	Economic Consequence
Submit more quality clips	Cubic scoring: 3 passing samples = 27x reward vs. 1
Maintain consistency	Failure gating: one bad interval = zero weight until recovery
Produce original content	Overlap pruning: duplicates reduce passing sample count
Write accurate captions	Semantic checks: fabricated or generic captions are rejected
Tag categories honestly	Category validation: misclassified content is rejected
Submit correct metadata	Hash/resolution checks: any mismatch rejects the full interval

6. Competitive Position

6.1 vs. Centralized Dataset Providers

Scale AI, Labelbox, and Appen produce high-quality training data at enterprise prices on enterprise timelines. Good work, if you can afford it and wait for it. NexisGen operates differently in three ways that matter:

- **Always running.** 24/7 production on a 20-minute interval cadence. No project scoping, no procurement cycles, no delivery delays.
- **Permissionless.** Anyone with a Bittensor wallet and an R2 account can mine. Quality determines earnings, not vendor relationships.
- **Provable verification.** Every clip is hash-verified, resolution-checked, and semantically validated. Centralized providers ask you to trust their QA. NexisGen proves it cryptographically.

6.2 vs. Open Dataset Projects

LAION, Common Crawl, WebVid, Common Voice, and academic datasets provide scale but not reliability. Quality control is manual and breaks at volume. Provenance is usually unclear. Deduplication is an afterthought. And each project covers one modality with no shared framework across data types.

Every sample in the NexisGen corpus passed a spec-defined validation pipeline. Every hash was verified. Every annotation was semantically checked against the actual content. Every category label was validated by a foundation model. There is no "trust the contributor" step. As new specs activate, the same guarantees extend to audio, image, and text without building new infrastructure.

6.3 vs. Other Bittensor Subnets

Aspect	Typical Data Subnets	NexisGen
Validation depth	Single-pass checks	Eight-phase pipeline with semantic and category verification
Deduplication	Per-miner only	Global cross-miner overlap index with timestamp arbitration
Quality incentive	Linear scoring	Cubic scoring with single-interval failure gating
Data provenance	Self-reported	SHA256-verified with source authentication
Storage model	Miner-hosted endpoints	Structured R2 with on-chain credential discovery
Sampling	Random or manual	Deterministic, seeded, reproducible across all validators

6.4 Honest Trade-Offs

Trade-Off	Mitigation
Owner-validator centralization at launch	Global overlap index and dataset curation will decentralize to validator consensus
YouTube-only source platform	Verifiable provenance justifies the constraint; additional platforms in Phase 2
External API dependency for semantic checks	Fail-open design prevents API outages from halting the network
Video-only at launch	Spec architecture is modality-agnostic; audio, image, and text on roadmap for 2027
Nature/landscape category only at launch	Spec system designed for multi-category expansion
Caption quality bounded by vision-model capability	Semantic verification catches the worst failures; model improvements lift the floor

7. Roadmap

Phase 1: Foundation (Current)

- Single spec: video_v1 - captioned 5-second clips at 1280x720
- Nature/landscape/scenery category focus
- YouTube-only source with verifiable provenance
- Full eight-phase validation pipeline with semantic and category checks

- Cubic scoring with single-interval failure gating
- On-chain credential discovery via Bittensor commitments
- Owner-validator maintains global overlap index

Phase 2: Video Scale (Q3-Q4 2026)

- Additional categories: action, dialogue, urban, wildlife, sports, tutorial
- Higher resolution specs: video_v2 at 1920x1080, video_v3 at 4K
- Variable duration: 10s, 30s, 60s clips
- Expanded source platforms beyond YouTube
- Decentralize overlap index to validator consensus
- Cross-category quality metrics and dashboards

Phase 3: Audio & Image Expansion (Q1-Q2 2027)

- audio_v1: speech transcription datasets with speaker identification, language tags, transcript verification
- audio_v2: music and environmental sound datasets with genre, instrument, and scene classification
- image_v1: captioned high-resolution stills with object detection annotations and scene classification
- Cross-modal alignment datasets: paired video-audio, image-text for retrieval training
- Modality-specific validation: audio quality checks (sample rate, noise floor), image checks (resolution, compression)

Phase 4: Text & Structured Data (Q3-Q4 2027)

- text_v1: curated text corpora - deduplicated, quality-filtered, domain-classified with provenance
- instruction_v1: instruction-response pairs with quality scoring and semantic verification
- Multi-clip narrative sequences for temporal reasoning training
- Community-contributed dataset specs via governance
- Cross-modality bundles: synchronized video + audio + text for multimodal training

Phase 5: Marketplace (2028+)

- Direct dataset access API for model trainers
- Per-modality and per-category quality leaderboards
- Custom spec requests from consumers
- Revenue sharing from dataset licensing

- Consumer payments + alpha token buyback-and-burn
- Provenance certificates for compliance and audit

8. Conclusion

Pay people for verified quality. Penalize them for garbage. See if the data gets better. That is what NexisGen tests.

Miners produce training data against a published spec. An eight-phase pipeline determines what passes. Cubic scoring makes the gap between good and mediocre brutal. Failure gating makes inconsistency expensive. TAO flows to whoever earned it. The same mechanism works whether the data is video, audio, images, or text.

The system launches with video because that is where the bottleneck bites hardest. The spec architecture is built to extend. The owner-validator is centralized at launch because the overlap index has to work before it can be distributed. These are sequencing decisions, not permanent constraints.

Whether NexisGen works depends on a question no whitepaper can answer: can decentralized miners, competing under protocol-enforced quality rules, produce training data good enough that labs would choose it over building internal pipelines? The validation framework, the scoring, the anti-gaming system - all of it exists to make that bar reachable. What happens next is up to the miners.

References

Market and Industry

Grand View Research - AI Training Dataset Market Size & Trends Analysis, 2025-2030

Epoch AI - Key Trends in AI Data (2025) - Data requirements for frontier model training

McKinsey - The Economic Potential of Generative AI - $2.6-4.4T annual economic impact estimate

Bittensor Ecosystem

Bittensor Network - Decentralized AI network protocol

Bittensor Whitepaper - Yuma Rao (2021)

Bittensor Documentation - Subnet registration, metagraph, weight submission

Training Datasets and Benchmarks

WebVid-10M - Bain et al. (2021) - Large-scale text-video dataset

HowTo100M - Miech et al. (2019) - Learning from instructional videos

InternVid - Wang et al. (2023) - Large-scale video-text dataset

LAION-5B - Schuhmann et al. (2022) - Large-scale image-text dataset with video extensions

Common Voice - Mozilla - Multilingual speech corpus

LibriSpeech - Panayotov et al. (2015) - Large-scale ASR corpus from audiobooks

The Pile - Gao et al. (2020) - Large-scale diverse text corpus for language modeling

Multimodal AI Systems

GPT-4 Technical Report - OpenAI (2023)

Gemini: A Family of Highly Capable Multimodal Models - Google DeepMind (2023)

GPT-4V(ision) System Card - OpenAI (2023)

Data Quality and Provenance

Data Provenance Initiative - Longpre et al. (2023) - Auditing training data for LLMs

Do Datasets Have Politics? - Peng et al. (2021) - Disciplinary values in dataset development

Documenting Data for People - Gebru et al. (2021) - Datasheets for datasets

Decentralized Data Networks

Ocean Protocol Whitepaper - Decentralized data exchange protocol

Filecoin Whitepaper - Decentralized storage network