
NATIVE MULTIMODALITY
VISUAL UNDERSTANDING
VISUAL UNDERSTANDING
AUDITORY PROCESSING
REASONING ENGINES
SYMBOLIC LOGIC
DECISION MAKING
DECISION MAKING
GOOGLE FOUNDATIONAL MODELS by Michaël BETTAN
A comprehensive structural blueprint of native multimodality, reasoning engines, and the specialized AI fleet.
SPECIALIZED AI FLEET
TASK-SPECIFIC MODELS
TASK-SPECIFIC MODELS
DOMAIN ADAPTATION
UNIFIED ARCHITECTURE
GLOBAL DEPLOYMENT
SCALABLE INFRASTRUCTURE
GLOBAL DEPLOYMENT

The Paradigm Shift: Native Multimodality

Legacy: Late Fusion

Native: Early Fusion

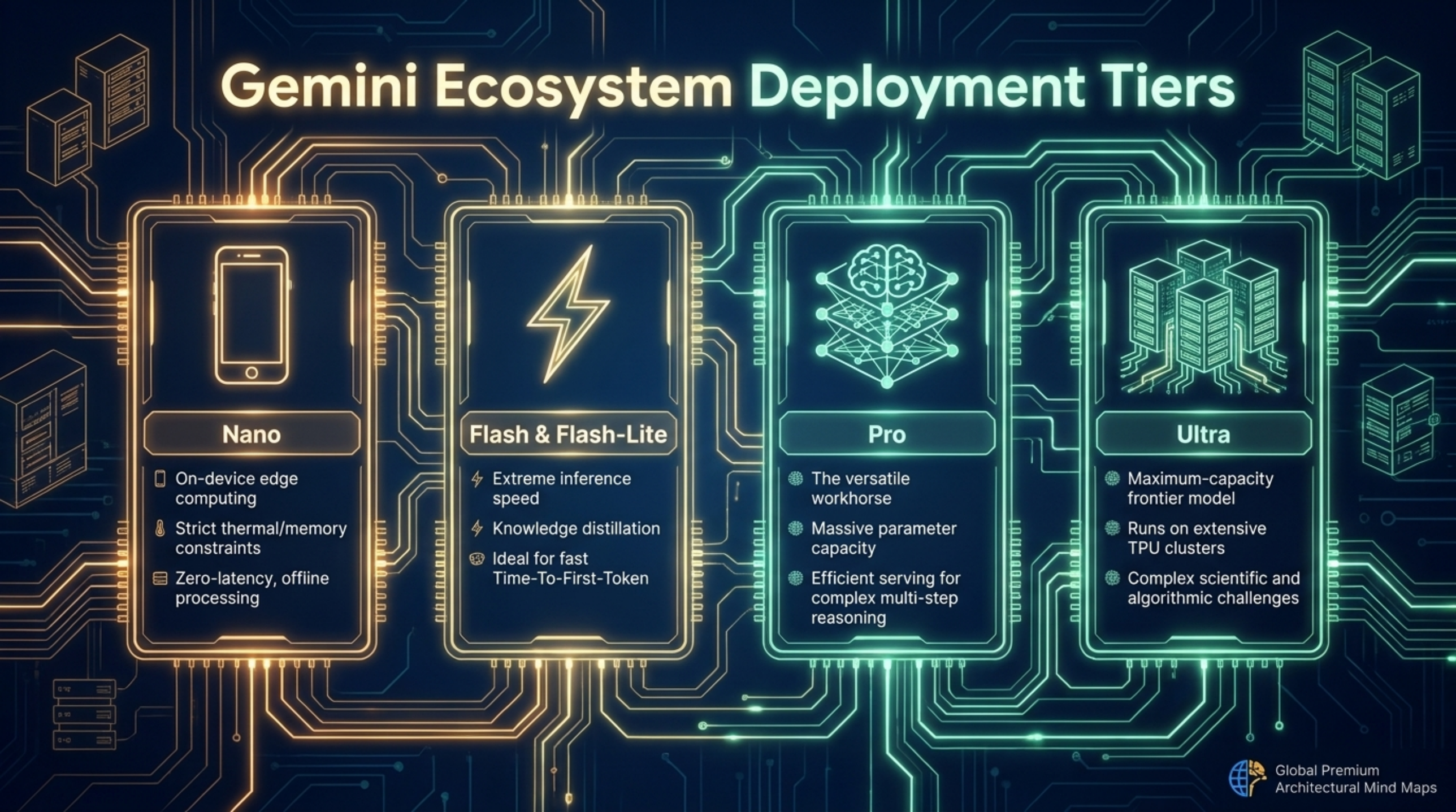
The Backbone: Sparse Mixture-of-Experts (MoE)

Moved from PaLM 2 (dense, text-first) to Gemini natively multimodal MoE routing.

Conditionally routes tokens to specialized expert subsets.

Decouples total capacity from per-token compute cost, enabling massive scale.

Gemini Ecosystem Deployment Tiers



Nano

- On-device edge computing
- Strict thermal/memory constraints
- Zero-latency, offline processing

Flash & Flash-Lite

- Extreme inference speed
- Knowledge distillation
- Ideal for fast Time-To-First-Token

Pro

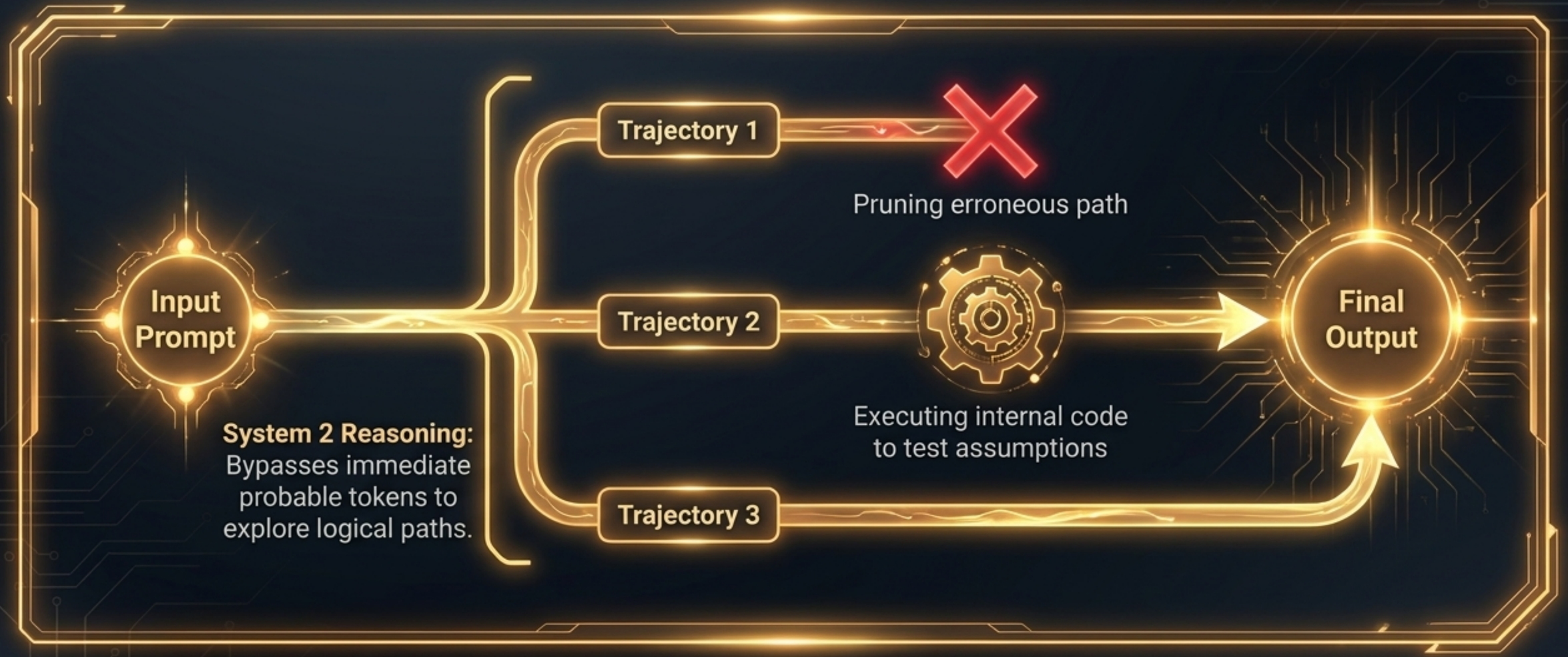
- The versatile workhorse
- Massive parameter capacity
- Efficient serving for complex multi-step reasoning

Ultra

- Maximum-capacity frontier model
- Runs on extensive TPU clusters
- Complex scientific and algorithmic challenges

Gemini 3.1 Pro: The Deep Think Paradigm

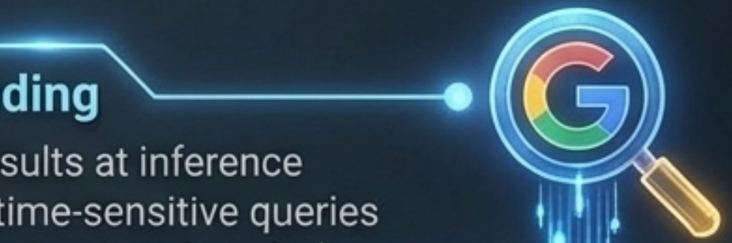
Apex reasoning model optimized for agentic workflows with controllable chain-of-thought.



Context Architecture & Live Grounding

Dynamic Search Grounding

- Injects real-time Search results at inference
- Reduces hallucination on time-sensitive queries
- Distinct from RAG (uses native Google Index)



Strict 65,536-token output limit

Tokenization Mechanics Across Modalities

Text

256,000-token vocabulary using SentencePiece unigram.

Visual

Dynamic tiling. Scaled/cropped to 768x768 tiles deterministically mapping to 258 tokens.

Video

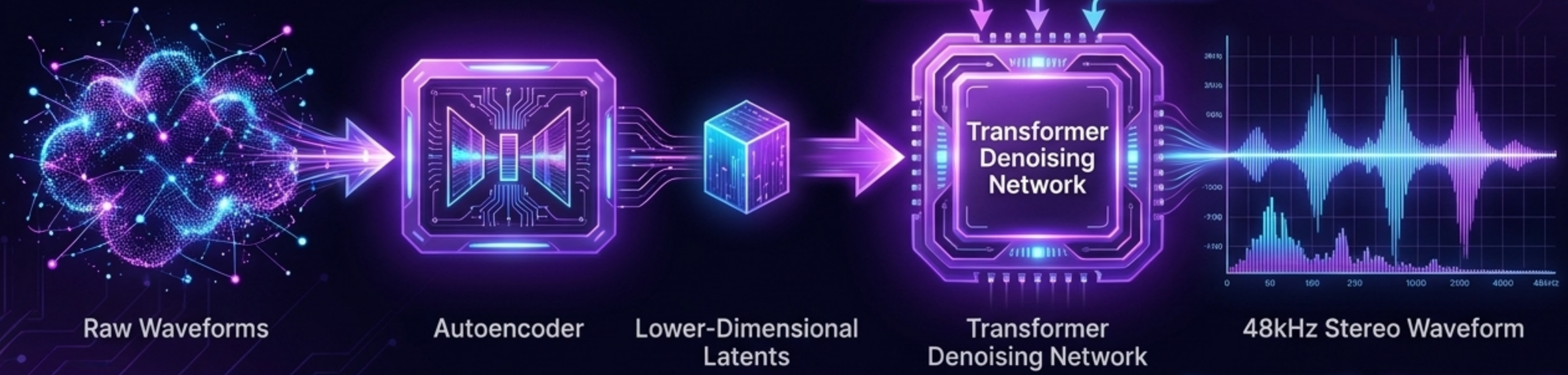
Parallel streams. Configurable sampling (1 FPS = ~263 tokens/sec). Compressible to 66-70 tokens for multi-hour context.

Timestamp Tokens: Anchors parallel streams for temporal coherence

Encoded continuously at an efficient fixed rate of 32 tokens per second.

Lyria 3: Temporal Latent Diffusion Audio

Abandons autoregressive prediction to eliminate statistical drift.



Lyria 3 Clip: 30s rapid prototyping

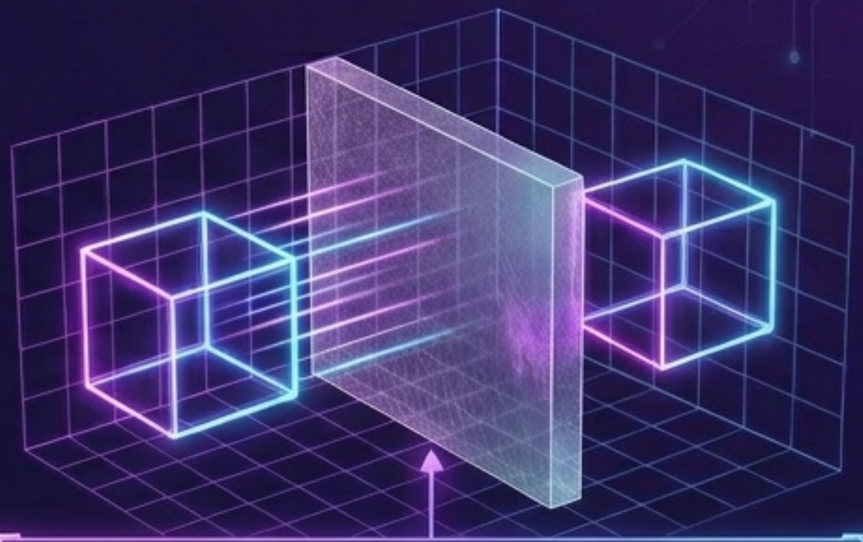
Lyria 3 Pro: 184s full compositions

Visual Generation: Veo 3.1 & Nano Banana 2

Joint Latent Diffusion

Emergent perfect synchronization (e.g., footsteps align perfectly with visual impact). No bolted-on audio track.

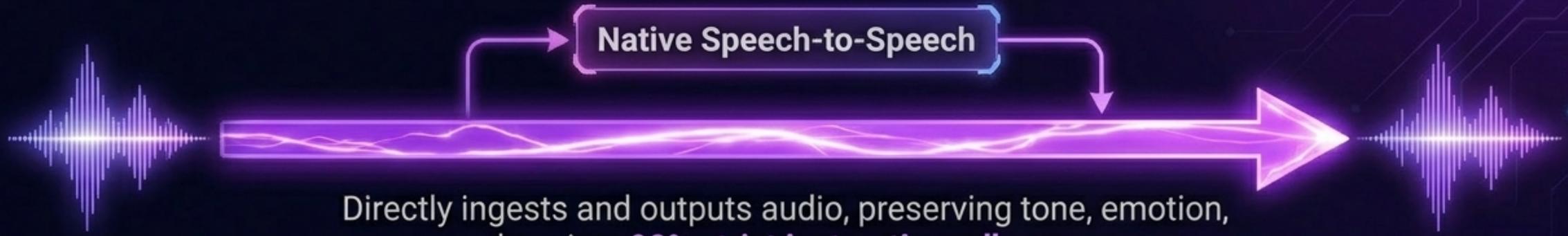
Spatial & Temporal Attention



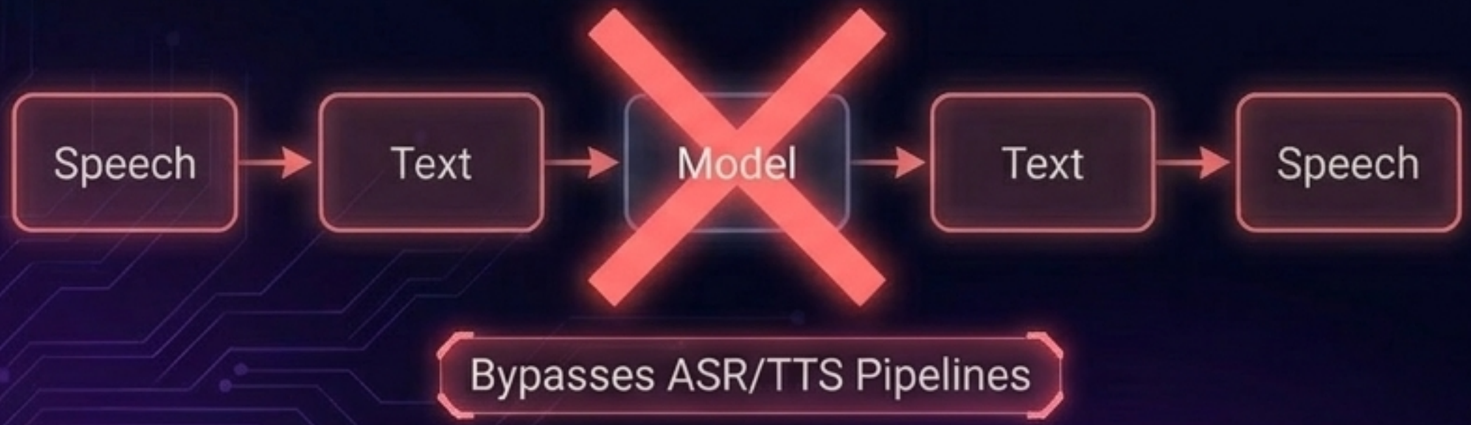
Maintains structural identity of up to 5 characters and 14 objects natively. Occluded objects reappear correctly without morphing.

131,072 context / Native PDF ingest / 4K Resolution

Gemini Audio: Pure Acoustic Intelligence



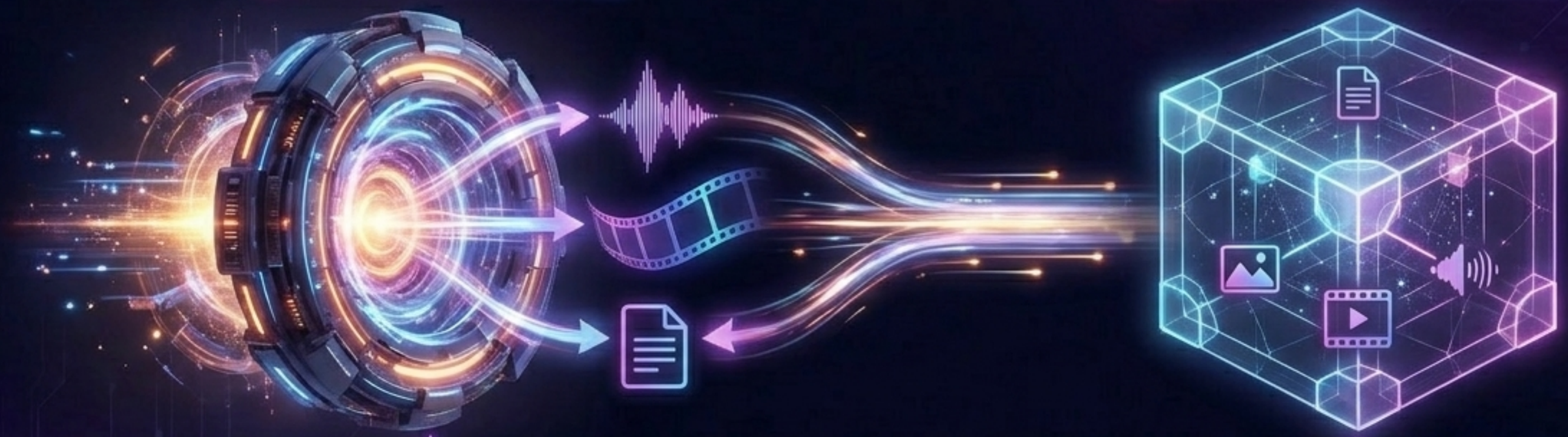
Directly ingests and outputs audio, preserving tone, emotion, and pacing. **90% strict instruction adherence.**



- **Live Speech Translation:** Native style transfer matching intonation and pitch.
- **Complex Tool Use:** 71.5% ComplexFuncBench Audio score.
- **Built-in Noise Filtering** for loud environments.

The Multimodal Live API & Unified Memory

WebSocket (WSS)



Gemini Embedding 2

Stateful, bidirectional streaming. Streams raw 16-bit PCM audio, video frames, and text simultaneously. Outputs raw 24kHz affective dialog.

Unified Vector Space. Maps all modalities into a single mathematical store. **Cross-modal retrieval:** Image query instantly retrieves video timestamps.

SynthID: Imperceptible Cryptographic Provenance

Embeds cryptographic signatures directly into data structures at inference.

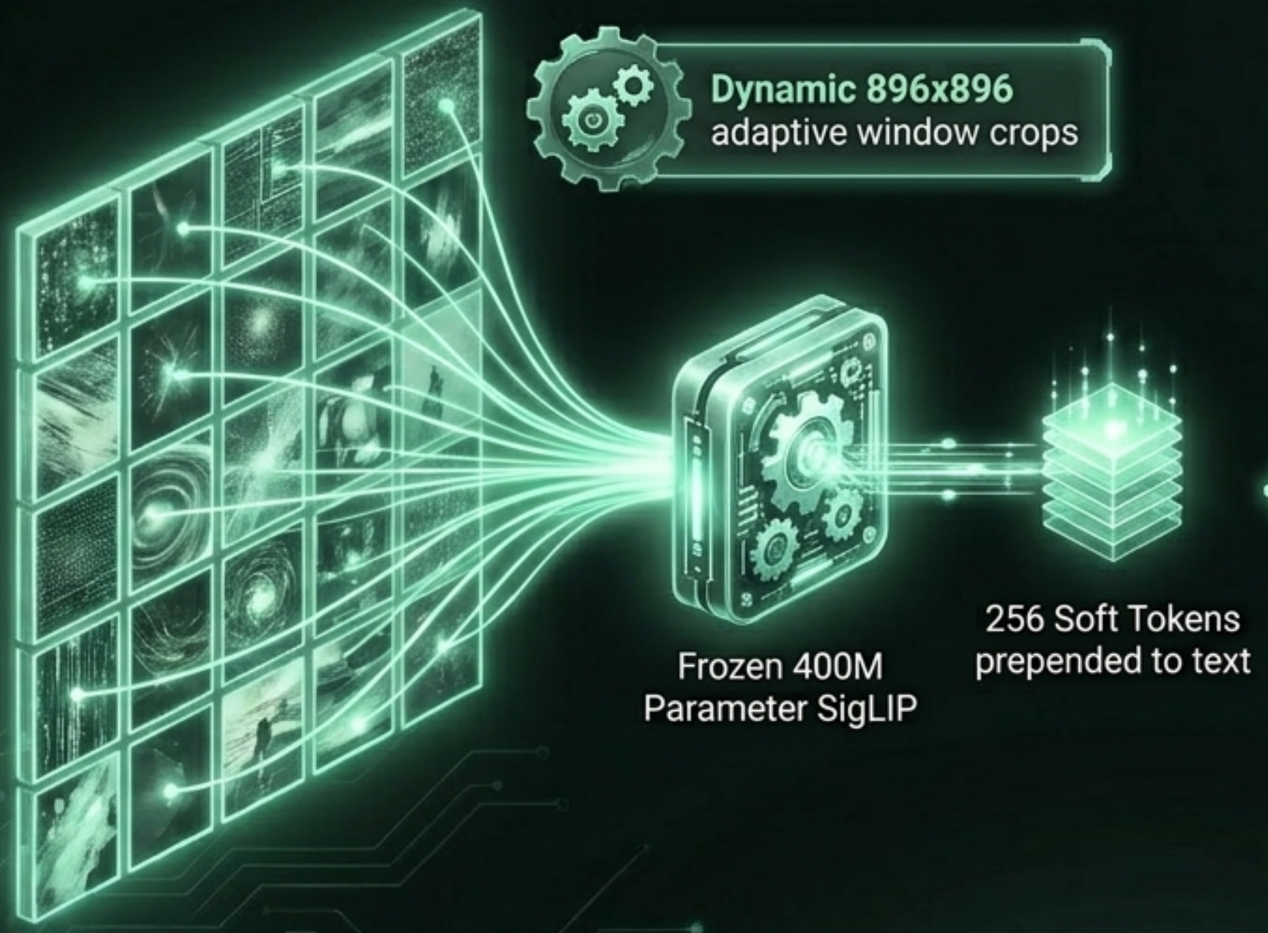
Tournament Sampling: Subtle skewing of probability distributions (logits) of word choices via pseudo-random g-function.

Psychoacoustic Spectral Embedding: Hides data in frequencies where human hearing is weakest. Survives the analog hole.

Pixel-Level Latent Manipulation: Distributed holistically across the spatio-temporal volume. Survives cropping and lossy compression.

Gemma 3: Intelligence Constrained

Bringing the Gemini core architecture to local edge environments.

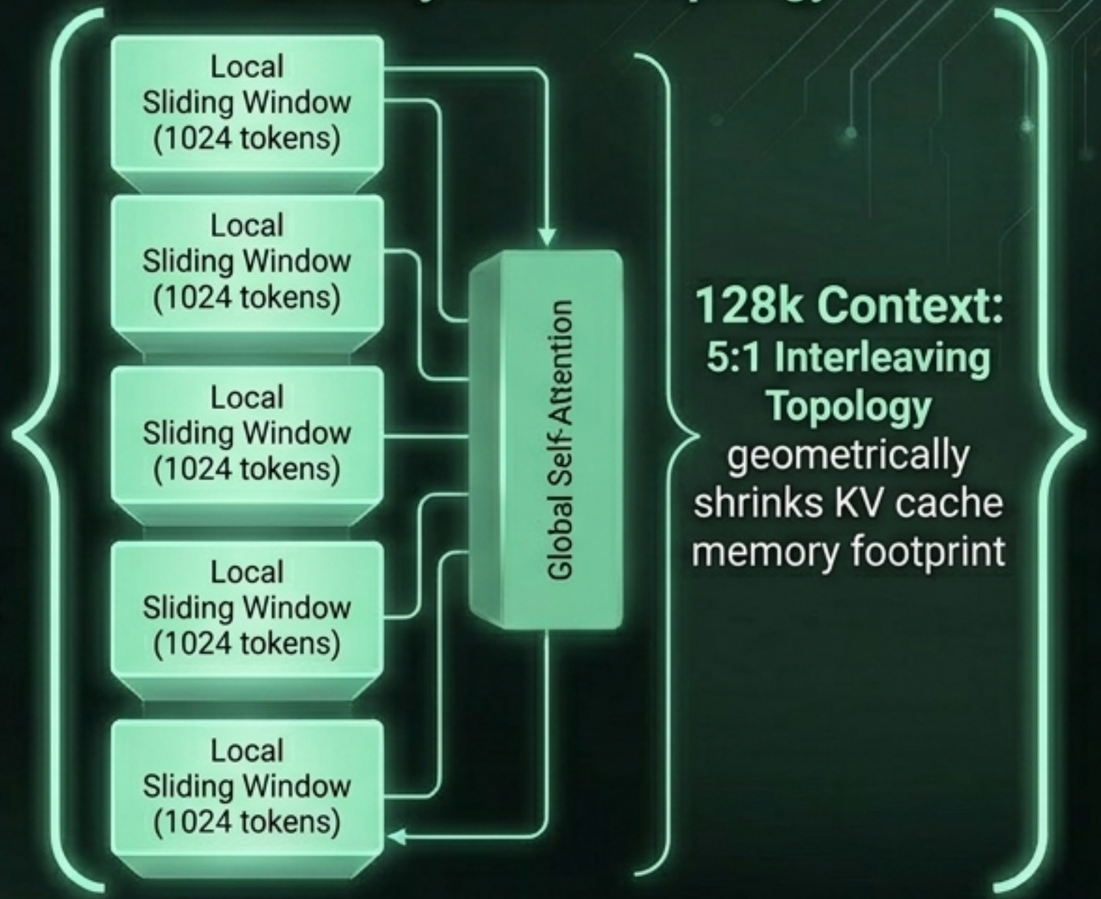


Dynamic 896x896
adaptive window crops

Frozen 400M
Parameter SigLIP

256 Soft Tokens
prepended to text

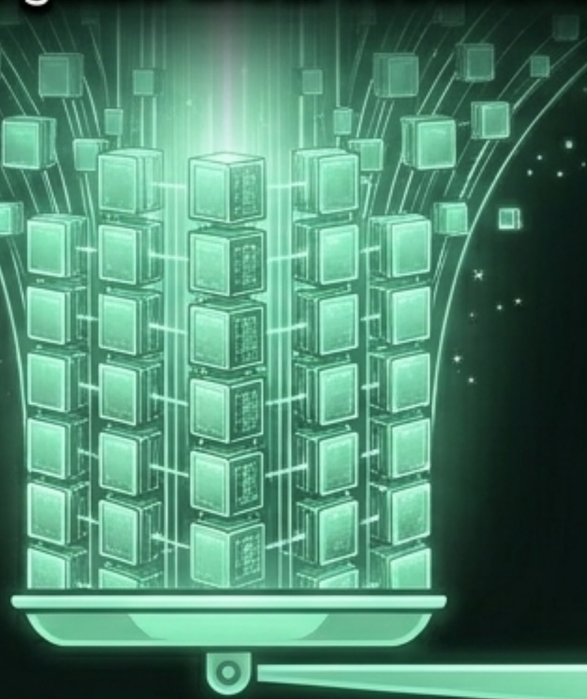
Memory Context Topology



128k Context:
5:1 Interleaving Topology
geometrically shrinks KV cache memory footprint

RecurrentGemma: The Griffin Architecture

Hybrid gated linear recurrence + local sliding window attention.



Expanding KV Cache
(Memory Exhaustion)

Traditional Attention

Fixed-Size
Internal State

RecurrentGemma

The Trade-Off: Accepts a slight drop in needle-in-a-haystack retrieval in exchange for substantially higher sampling throughput and drastically lower memory consumption for long sequences at the edge.

AlphaFold 3: Generative Molecular Simulation

Chaotic Cloud of Noise Atoms

Extreme Granularity: 1 token per atom.

Conditional Diffusion Module

Physically Stable 3D Molecular Complex

The Paradigm Shift: Replaces the Evoformer with the Pairformer, focusing directly on pairwise atom interactions rather than Multiple Sequence Alignments.

AlphaGeometry 2: Neuro-Symbolic Logic

84%
IMO Geometry Success Rate

Engine 1: Neural Language Model (Gemini)

- Heuristic & Intuitive
- Predicts mathematically probable auxiliary constructs (e.g., drawing unstated lines).

Engine 2: Symbolic Engine (DDAR)

- Rigid & Infallible
- Computes thousands of exhaustive rule-bound deductions to verify the proof.

The Specialist Fleet: Biology & Pattern Decoders

AlphaProteo

Designs novel, high-affinity protein binders for diverse target proteins to accelerate drug discovery.

AlphaMissense

Classifies missense mutations to identify pathogenic variants within the human genome.

DolphinGemma

Specialized decoding model engineered to parse and model complex animal communication patterns.

Aeneas

Neural sequence modeling applied to the humanities; restores missing text and dates damaged ancient inscriptions.

The Specialist Fleet: Math, Code & Environment

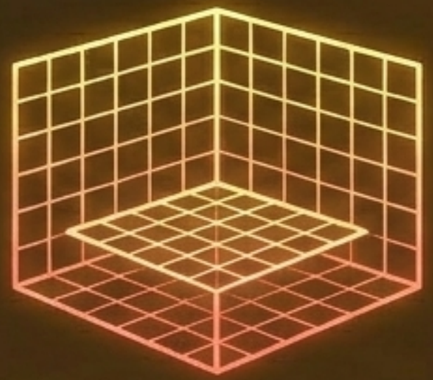
Algorithms & Code

AlphaTensor: Discovers computationally efficient matrix multiplication algorithms.

AlphaDev: Discovers faster assembly-level sorting/hashing via reinforcement learning.

AlphaCode: Generates competitive-level code solutions via mass sampling.

Spatial & World Models



Genie 2 / 3: Generates interactive 2D/3D environments from prompts.

SIMA 2: Collaborative generalist agent reasoning wirtual 3D worlds.

Climate & Systems

GraphCast: High-resolution probabilistic global and extreme weather forecasting.

Synthesis: The Foundational Ecosystem Blueprint