Subquadratic SubQ: An LLM That Reads 12 Million Tokens in One Prompt

Kabui, Charles

2026-05-28

Read at ToKnow.ai

On May 5, 2026, Miami startup Subquadratic launched SubQ with $29 million in seed funding. The model accepts 12 million tokens in a single prompt: roughly nine million words, or about six months of pull requests against the React codebase. Standard transformer attention compares every token to every other token, so cost scales quadratically. SubQ's Subquadratic Selective Attention (SSA) picks which positions actually matter, and the selection step itself

scales linearly. Reported scores: 95.6% on RULER at 128K, 86.2% on MRCR v2 (8-needle) at 1M tokens against GPT-5.5 at 74.0% and Claude Opus 4.7 at 32.2%, 81.8% on SWE-Bench Verified, and 92.1% on needle-in-a-haystack retrieval at the full 12M context.

Most frontier models cap at 1 million tokens because the cost becomes prohibitive past that point. SubQ claims 52x faster inference than dense attention at 1M tokens and nearly 1,000x less compute at 12M, priced at roughly one-fifth of Claude Opus or GPT-5.5. An independent Appen audit measured a 56x wall-clock speedup over FlashAttention-2 at 1M tokens on B200 hardware (381 ms vs. 21.4 seconds). The practical change: a developer can drop an entire codebase, its pull requests, and its issue history into one API call instead of writing retrieval pipelines.

Magic.dev announced a 100-million-token model in August 2024 with similar "1,000x efficiency" claims and raised more than $500 million; nothing public has shipped since. SubQ is in private beta, weights are closed, and each benchmark ran once. The architecture is plausible and the speed numbers are externally verified, but the long-context category has burned investors before.

Read More: Anthropic's Claude Sonnet 4.6 with 1 million token context showed how labs were pushing the 1M-token frontier with curation rather than architectural change.

Sources:

- Subquadratic: Introducing SubQ (May 5, 2026)
- SiliconANGLE: Subquadratic launches with $29M to bring 12M-token context windows to AI (May 5, 2026)
- The New Stack: The context window has been shattered (May 5, 2026)
- Appen: Independent benchmark of Subquadratic's SSA kernel (May 11, 2026)
- Subquadratic homepage
