VOLTROPY

Vast-10M

Technical Report

Clint Ehrlich
Voltropy PBC
clint@voltropy.com

Ted Blackman
Voltropy PBC
ted@voltropy.com

September 29, 2026

Abstract

We introduce **Vast-10M**, a family of LLMs with ten million tokens of native context. The key enabling technology is a new form of attention, **Voltropy Scalable Attention (VSA)**, which can be retrofitted to existing transformer models.

Prior methods for extending context beyond a million tokens significantly degraded model intelligence. Our results indicate that VSA not only eliminates this tradeoff but improves reasoning quality at shorter contexts.

We ablate VSA by testing base DeepSeek and GLM models against their Vast-10M counterparts on the BEAM (Beyond a Million Tokens) benchmark. Every VSA-augmented model outperforms its corresponding base model at 100k, 500k, and 1M tokens. The flagship model, Vast-10M-Flash, achieves a higher score at 10M tokens than its base model achieves at 1M tokens. At 1M tokens, it beats Anthropic's Fable 5.1 and is on par with OpenAI's GPT-6 Astra.

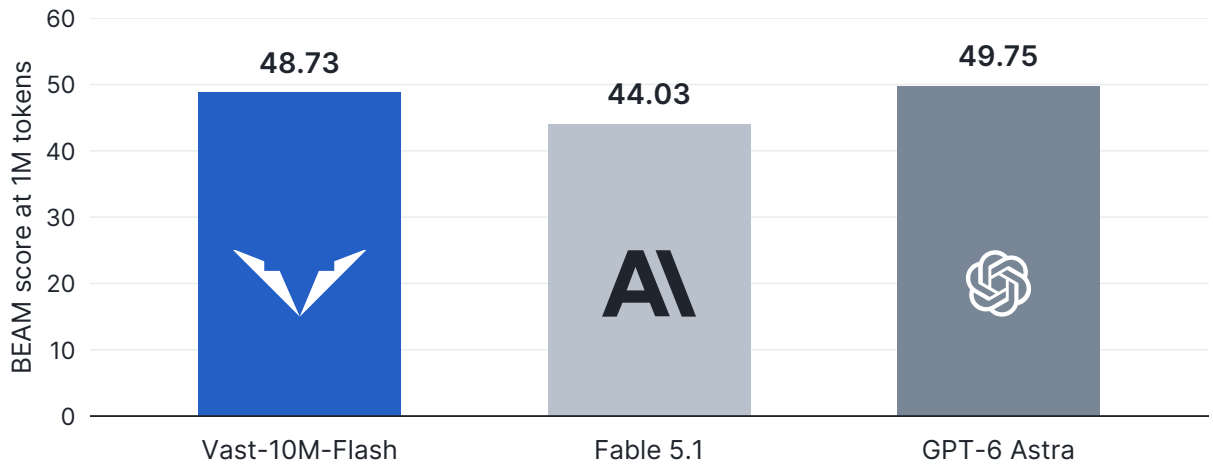

Figure 1. Vast-10M-Flash surpasses Fable 5.1 on the 1M BEAM tier and achieves parity with GPT-6 Astra.

Table 1. BEAM scores by nominal context tier (0–100; higher is better).

Model	100k	500k	1M	10M
Vast-10M-Flash (VSA)	52.66	50.22	**48.73**	**40.20**
DeepSeek V4.0 Flash	43.71	44.72	39.69	—
Vast-10M-Pro (VSA)	43.93	44.50	43.76	**36.09**
DeepSeek V4.0 Pro	42.28	41.05	38.02	—
Vast-10M-Medium (VSA)	49.02	44.75	45.67	**32.74**
GLM-5.2	39.50	40.49	40.06	—
DeepSeek V4.1 Flash	45.47	44.94	46.00	—
GPT-6 Astra	55.56	56.15	49.75	—
Claude 5.1 Fable	60.98	57.26	44.03	—

Table 2. VSA gains in BEAM score points, with relative increases over each base model in parentheses.

Model conversion	100K	500K	1M
DeepSeek V4.0 Flash → Vast-10M-Flash	**+8.95** (+20.48%)	**+5.50** (+12.30%)	**+9.04** (+22.77%)
DeepSeek V4.0 Pro → Vast-10M-Pro	**+1.66** (+3.92%)	**+3.45** (+8.40%)	**+5.74** (+15.10%)
GLM-5.2 → Vast-10M-Medium	**+9.52** (+24.09%)	**+4.26** (+10.53%)	**+5.61** (+14.00%)

1 Vast-10M and VSA

Prior methods for extending LLM context windows have required compromising model intelligence, either by approximating dense attention [1] in ways that degraded the quality of shorter-distance reasoning or by abandoning the transformer architecture altogether, often in favor of recurrent architectures like State Space Models [2]. We sought to eliminate this tradeoff by developing a new form of attention, Voltropy Scalable Attention (VSA), with novel scaling properties. VSA is designed to expand the context windows of legacy transformer models by an order of magnitude while simultaneously improving their reasoning at shorter contexts.

To showcase VSA, we replaced the attention mechanisms in three leading open-source models with VSA to create the Vast-10M family of models, each of which has ten million tokens of native context. Vast-10M-Flash, our flagship model, is based on DeepSeek-v4.0-Flash [3]. To demonstrate VSA's modularity across model architectures and parameter counts, we also tested Vast-10M-Pro, based on DeepSeek-v4.0-Pro [4], and Vast-10M-Medium, based on GLM-5.2 [5].

We benchmarked both the original models and the VSA versions on the BEAM (Beyond a Million Tokens) eval [6]. BEAM was developed to evaluate long-term memory in LLMs, including retrieval-augmented generation (RAG) and other memory systems; its ten-million-token tier exceeds the native context windows of the baselines evaluated in the original paper. For our purposes, the fact that BEAM tests reasoning from 100k to 10M tokens was ideal, because it meant that we could investigate whether VSA boosted answer quality at shorter contexts and, if so, how much of that improvement was retained as context grew by two orders of magnitude.[1]

2 BEAM Results

Our testing indicates that VSA has two novel properties that distinguish it from prior forms of transformer attention. Table 1 reports the complete scores from our evaluation; Appendix A describes the evaluation procedure and limitations.

2.1 Multi-Scale Reasoning Improvements

VSA did not simply extend the maximum length of the context window, but materially improved reasoning quality at all shared context lengths tested. This effect was dramatic and consistent: every VSA-equipped model beat its non-VSA counterpart at every shared context length. Gains ranged from 1.66 to 9.52 BEAM points across the nine paired comparisons (Figure 2 and Table 2).

[1] We chose to discontinue use of the shorter OOLONG benchmark [7] because our prior testing uncovered parametric contamination [8, Section 5].

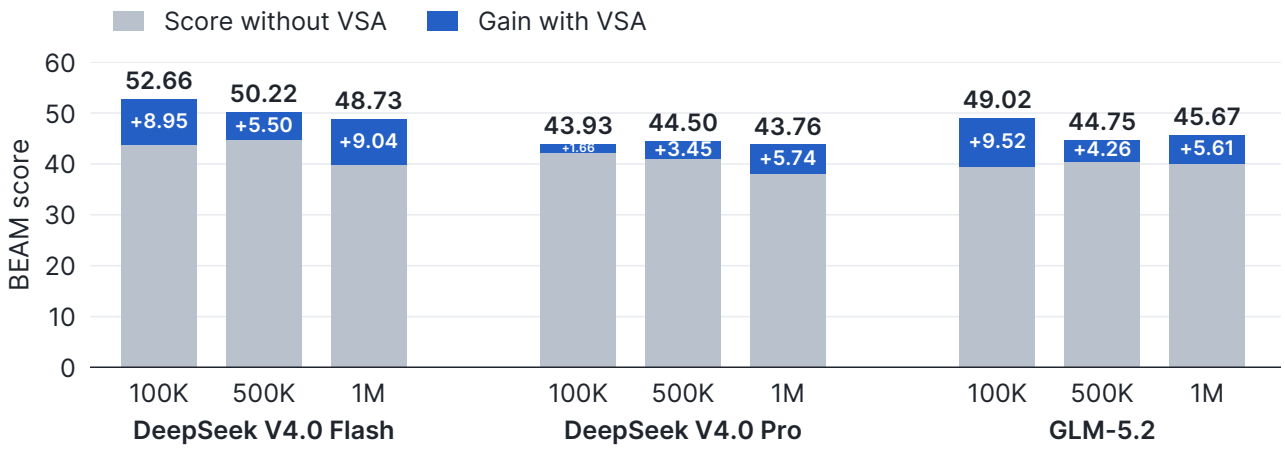

Figure 2. VSA improves BEAM scores for all models at all context lengths.

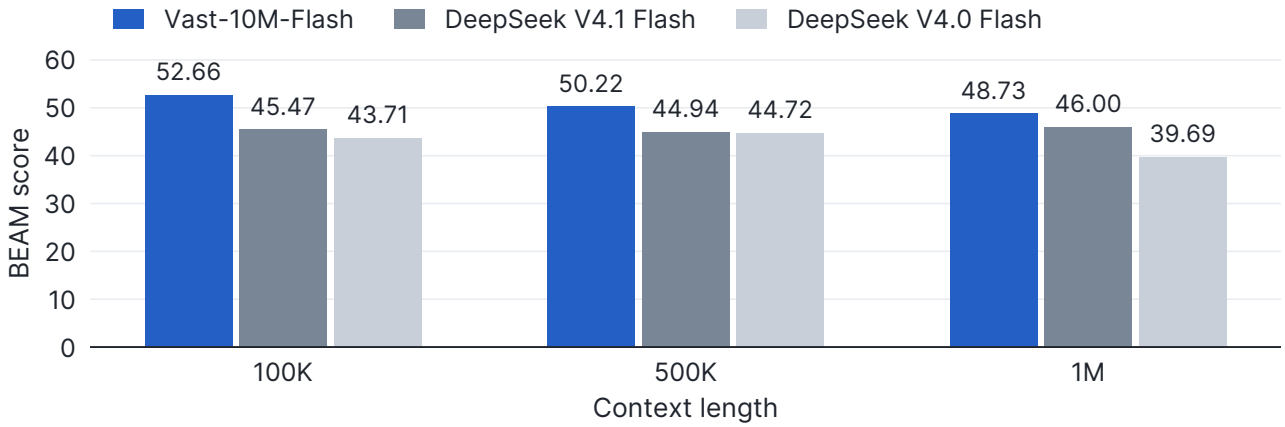

Figure 3. Vast-10M-Flash achieves higher BEAM scores than both its base model and its base model's successor.

Vast-10M-Flash outperformed not only its base model, DeepSeek-v4.0-Flash, but also that model's successor, DeepSeek-v4.1-Flash, at every shared context length (Figure 3). The v4.1 model includes a number of architectural innovations, including CSA2 attention, an encoder/decoder split, and Engram memory [9]. Yet adapting a legacy model to use VSA yielded a larger improvement in long-context reasoning than training a new model with a state-of-the-art architecture. Indeed, the v4.0 model with VSA scored higher at 1M tokens (48.73) than the non-VSA v4.1 model scored at 100k tokens (45.47).

This does not reflect a particularly low score from DeepSeek's v4.1, but rather an extremely high score from Vast-10M-Flash, as evidenced by comparisons to the strongest closed-source models in the world. Vast-10M-Flash's score at 1M tokens (48.73) was higher than Anthropic's Claude 5.1 Fable [10] (44.03) by 4.70 points (10.68%). Vast-10M-Flash achieved approximate parity with OpenAI's GPT-6 Astra [11] (49.75) at 1M tokens, delivering 97.95% of its score, despite the v4.0-Flash base model (39.69) not being competitive with either the Anthropic or OpenAI flagships (Figure 1).

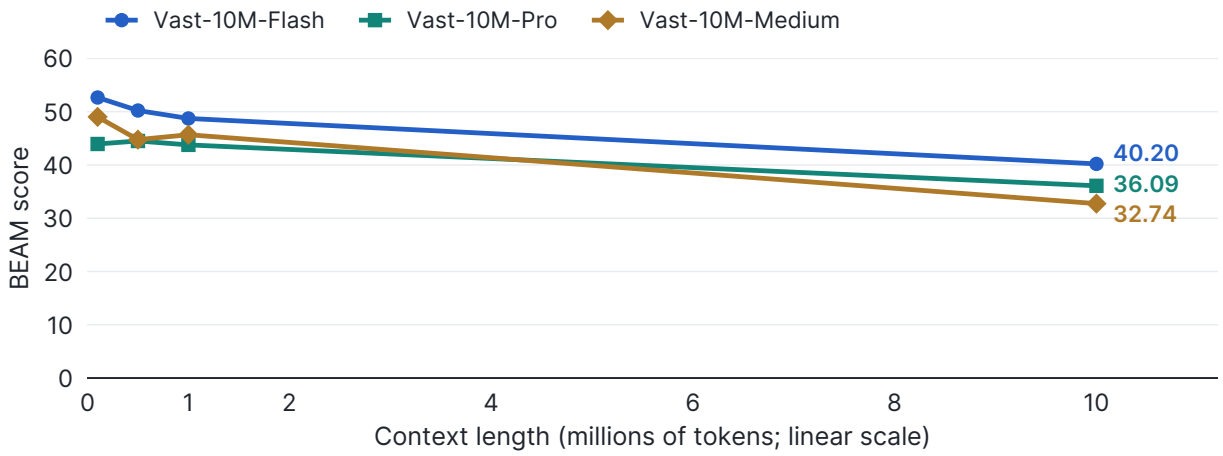

Figure 4. All Vast-10M models retain effective reasoning from 100k to 10M tokens.

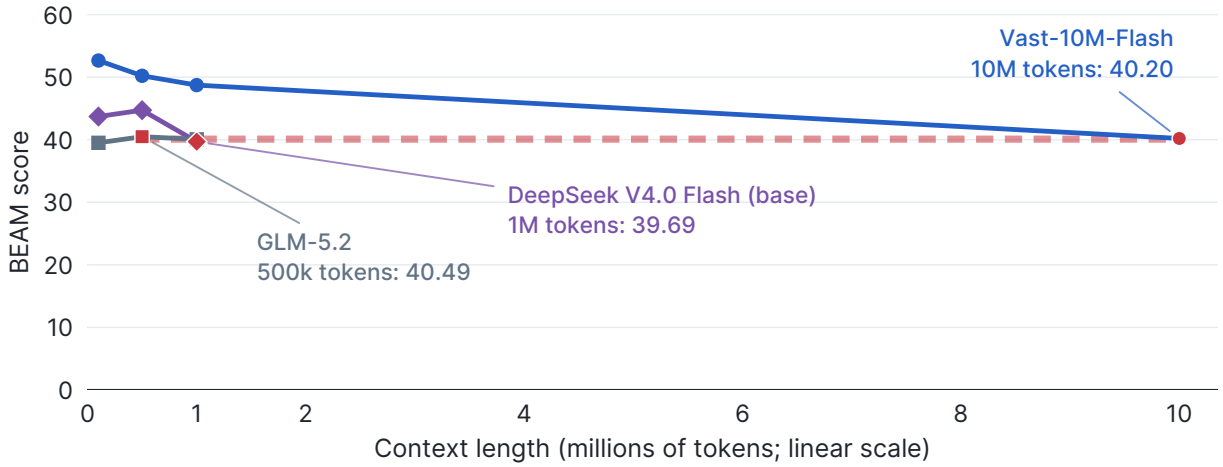

Figure 5. Vast-10M-Flash's BEAM score at 10M tokens matches DeepSeek v4.0 Flash's score at 1M tokens and GLM-5.2's score at 500k tokens.

2.2 Capability Preservation at Extreme Contexts

In addition to their improvements at shorter contexts, models equipped with VSA were able to reason at ten million tokens without using external memory systems. Our BEAM results indicate that for Vast-10M ten million tokens is not merely a nominal context window but a true functional cognitive workspace.

The flagship VSA-augmented model, Vast-10M-Flash, retained 82.49% of its 1M-token BEAM score at 10M tokens (40.20 versus 48.73). A two-order-of-magnitude increase in context size from 100k to 10M tokens produced only a 23.67% relative reduction in BEAM score, from 52.66 to 40.20 (Figure 4). To our knowledge, Vast-10M-Flash's score of 40.20 is the highest score ever achieved by a raw language model on the 10M-token BEAM tier.

Vast-10M-Flash scored higher at 10M tokens (40.20) than its base model DeepSeek-v4.0-Flash scored at 1M tokens (39.69). Its 10M-token score was effectively identical to GLM-5.2's score at 500k tokens (40.49). Vast-10M-Flash achieved 101.27% of its base model's score at ten times the context length and 99.29% of GLM-5.2's score at twenty times the context length (Figure 5).

Though lower, the other Vast-10M model scores at 10M tokens also appear to exceed previously reported BEAM scores for unassisted LLMs. Vast-10M-Medium and Vast-10M-Pro achieve unassisted scores of 32.74 and 36.09, respectively, at 10M tokens—higher than the RAG-assisted models in the original BEAM paper, including its strongest RAG baseline (24.90) and the highest LIGHT result in its main comparison (26.60) [6, Table 1].

3 Conclusion

Our BEAM testing demonstrates that VSA unlocks unprecedented long-context reasoning capabilities, even when retrofitted to existing transformer models. Across three model sizes and two model families, VSA consistently (1) improved reasoning quality across context lengths and (2) allowed models to reason effectively at ten million tokens without RAG or external memory. The increase in capability for the flagship model, Vast-10M-Flash, was large enough that it dominates its own base model's successor and rivals the strongest closed models in the world.

Appendix A

Technical Details and Limitations

All results were obtained according to the BEAM specification at BEAM Git revision `3e12035532eb85768f1a7cd779832b650c4b2ef9` [12]. These were full tests: all 400 questions at 100k tokens, all 700 questions at 500k, all 700 questions at 1M, and all 200 questions at 10M. The official LLM judge (`gpt-4.1-mini`) was used to judge all questions, with temperature set to 0. Results were scored according to BEAM's official formula [12].

Comparing BEAM scores between tiers provides a powerful heuristic for capturing differences in model capability, but the comparisons do not measure identical tasks, because the higher BEAM tiers include additional conversations and questions. However, the difficulty of the tasks should be comparable due to the generation procedure employed in the BEAM benchmark [6, Sections 2.2–2.3 and Appendix B], which was designed to generate conversations and questions of equivalent difficulty. To the extent that there may be some variation, it does not materially affect our results, due to the magnitude of the differences captured between context lengths (e.g., equivalent scores at 500k and 10M tokens).

We tested models only on the BEAM tiers that corresponded to their official context windows, e.g., Fable 5.1 and GPT-6 Astra were tested on the 1M and lower BEAM tiers, because OpenAI and Anthropic advertise that each of them has at least one million tokens of context [10, 11]. However, the tokenization scheme for a particular model may reduce the effective size of its context window below the length of the nominal tokens in the corresponding BEAM tier. When this occurred, we followed the official BEAM procedure of truncating the minimum input prefix required for the model to accept the BEAM conversation as input [12]. This procedure matches real-world usage: a user who attempts to fit 1M tokens of nominal context into a model whose 1M-token context window cannot accommodate the input will also have to truncate the conversation.

One confounding variable when measuring the integration of a new attention mechanism into an existing transformer model is the effect of whatever additional training is required to make the legacy weights compatible with the new attention mechanism. The greater the amount of retraining required, the greater the risk that an observed benchmark improvement is an artifact caused by what is effectively a fine-tune of the original model on the retraining corpora. If an attention mechanism requires extensive retraining, a robust ablation could require fine-tuning the original checkpoint with the legacy attention mechanism on the same corpora used to integrate the new attention mechanism. That was not necessary to obtain reliable ablations here, because VSA did not require extensive retraining for the selected model families.

References

1 Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer, 2020. URL https://arxiv.org/abs/2004.05150. arXiv:2004.05150.

2 Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces, 2023. URL https://arxiv.org/abs/2312.00752. arXiv:2312.00752.

3 DeepSeek-AI. DeepSeek-V4-Flash-0731, 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash-0731. Model card.

4 DeepSeek-AI. DeepSeek-V4-Pro, 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro. Model card.

5 Z.ai. GLM-5.2, 2026. URL https://huggingface.co/zai-org/GLM-5.2. Model card.

6 Mohammad Tavakoli, Alireza Salemi, Carrie Ye, Mohamed Abdalla, Hamed Zamani, and J. Ross Mitchell. Beyond a Million Tokens: Benchmarking and Enhancing Long-Term Memory in LLMs, 2025. URL https://arxiv.org/abs/2510.27246. arXiv:2510.27246; ICLR 2026.

7 Amanda Bertsch, Adithya Pratapa, Teruko Mitamura, Graham Neubig, and Matthew R. Gormley. Oolong: Evaluating Long Context Reasoning and Aggregation Capabilities, 2025. URL https://arxiv.org/abs/2511.02817. arXiv:2511.02817.

8 Clint Ehrlich and Theodore Blackman. LCM: Lossless Context Management, 2026. URL https://arxiv.org/abs/2605.04050. arXiv:2605.04050, Section 5.

9 DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the Limits of KV Cache Compression, 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash. Model card.

10 Anthropic. Claude Fable 5.1, 2026. URL https://platform.claude.com/docs/en/models/fable-5-1/whats-new-fable-5-1. Model documentation.

11 OpenAI. GPT-6 Astra, 2026. URL https://developers.openai.com/api/docs/models/gpt-6-astra. Model documentation.

12 Mohammad Tavakoli et al. BEAM: Official Code and Evaluation Implementation, 2026. URL https://github.com/mohammadtavakoli78/BEAM/tree/3e12035532eb85768f1a7cd779832b650c4b2ef9. Pinned revision 3e12035532eb85768f1a7cd779832b650c4b2ef9.