INDEPENDENT BENCHMARK EVALUATION

Appen

Subquadratic

Model Performance Evaluation to SubQ 1.1 Small Preview Performance Evaluation

An independent third-party benchmark assessment conducted by Appen Ltd.

APPEN.COM

Published: June 16, 2026
Edition: Round 2 (Public Brief)

Evaluated by: Jeanine Sinanan-Singh & Tahseen Rabbani, Appen Ltd.

Benchmarks: NIAH (1M to 12M), LiveCodeBench

About This Evaluation

Appen is a global leader in AI data services and model evaluation, with over 25 years of experience building, testing, and benchmarking AI systems across the world's most demanding applications. Appen was engaged by Subquadratic to independently assess the performance of their latest preview models across a suite of publicly recognised benchmarks. This brief is the public summary of Round 2; the full technical report is available on request.

Independence Statement

The Appen benchmark team operated with full independence throughout this assessment. Access was scoped exclusively to Subquadratic's API endpoints and authentication keys: no model weights, training datasets, fine-tuning configurations, or benchmark ground-truth labels were provided in advance. All results reflect authentic, uninfluenced model performance.

Key Results

Subquadratic's preview models sustain near-perfect long-context retrieval out to 12 million tokens and reach 89.7% pass@4 on LiveCodeBench.

100%
NIAH Retrieval
1M to 12M token context

89.7%
LiveCodeBench pass@4
78.0% pass@1 across 1,055 problems

12M
Verified Context Length
single-needle retrieval, 50 of 50

Full Results Summary

Benchmark	Metric	Result
NIAH (niah_single_1, RULER)	Retrieval accuracy, 1M	100%
NIAH (niah_single_1, RULER)	Retrieval accuracy, 2M	100%
NIAH (niah_single_1, RULER)	Retrieval accuracy, 6M	98%
NIAH (niah_single_1, RULER)	Retrieval accuracy, 12M	100%
LiveCodeBench	pass@1 (1,055 problems)	78.0%
LiveCodeBench	pass@4	**89.7%**

Long-Context Retrieval: Needle-in-a-Haystack

Needle-in-a-Haystack (NIAH) evaluates whether a model can locate a specific fact, the needle, embedded at varying depths within a very long context. The evaluation uses the niah_single_1 task (single needle) from the RULER suite[1], with 50 samples per context tier at temperature 0 and zero execution errors. Retrieval accuracy is defined as the target value appearing in the response.

> *Across 1M to 12M token contexts, the model retrieves the target value in 98 to 100% of samples, sustaining reliable retrieval at context lengths well beyond the reach of dense-attention models.*

At the 1M and 2M tiers the model returned the target value verbatim within a complete sentence on every sample (50 of 50), so retrieval and exact-match are both 100%. The 6M and 12M tiers, run on the longer-context configuration, hold at 98% exact-match.

Context tier	Model	Samples	Retrieval accuracy	Exact match
1M	subq-2m-preview-small	50	**100%**	100%
2M	subq-2m-preview-small	50	**100%**	100%
6M	subq-12m-preview-nano	50	**98%**	98%
12M	subq-12m-preview-nano	50	**100%**	98%

Code Generation: LiveCodeBench

LiveCodeBench evaluates code generation on competitive-programming problems collected continuously from contest platforms, with release-date filtering to limit contamination[2]. Appen evaluated 1,055 problems with four completions each (4,220 total), reporting pass@1 and pass@4 overall and by difficulty. pass@k is the fraction of problems with at least one passing completion in k samples.

Subset	Problems	pass@1	pass@4
Overall	1,055	78.0%	**89.7%**
Easy	322	92.9%	99.1%
Medium	383	85.7%	95.8%
Hard	350	55.9%	74.3%

Evaluated on subq-2m-preview-small.

[1] Hsieh, C.-P., Sun, S., Kriman, S., et al. (2024). RULER: What's the Real Context Size of Your Long-Context Language Models? COLM 2024. arXiv:2404.06654

[2] Jain, N., Han, K., Gu, A., et al. (2024). LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. arXiv:2403.07974

About Appen

Appen is a global leader in AI data collection, annotation, and model evaluation. With more than 25 years of experience and a network of over one million skilled contributors worldwide, Appen helps the world's leading technology companies develop safe, accurate, and high-performing AI models. Appen's evaluation services provide independent, rigorous third-party assessments trusted by enterprises and AI developers globally.

To learn more about Appen's AI evaluation and benchmarking services, visit appen.com.

Evaluation Team

Name	Title	Organization	Contact
Jeanine Sinanan-Singh	Director of GenAI Research	Appen Ltd.	jsinanansingh@appen.com
Tahseen Rabbani	Forward Deployed Engineer	Appen Ltd.	trabbani@appen.com

Full Report Available on Request

A more exhaustive technical report, including detailed methodology, per-sample results, model configurations, latency and error analysis, and signed attestations, is available to qualified parties upon request. This extended report is designated confidential and subject to Appen's standard non-disclosure terms. To request access, please contact the evaluation team directly at the addresses listed above.

Appen