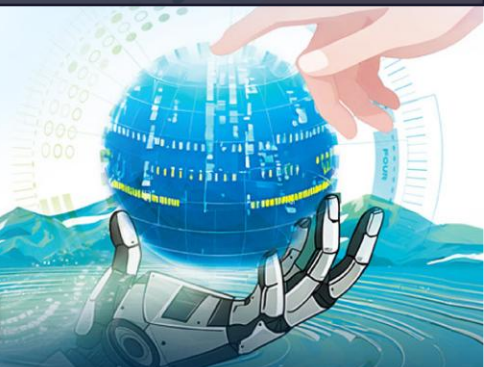
IEEE SMC 2026
Bellevue, WA USA



NeuroGenLLM: A Fast Spiking Neural Architecture Generation Framework via LLM-Guided Search

Yue Zhong[1], Zhijie Yang[2], Chao Xiao[1], Renzhi Chen[2], Shasha Guo[2], Rui Gong[1], Mingche Lai[1], and Lei Wang[2]*

[1] College of Computer Science and Technology, National University of Defense Technology, Changsha, China
[2] Defense Innovation Institute, Academy of Military Sciences, Beijing, China

国防科技大学
NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY

Motivation and key idea

SNN architecture search remains challenging

- SNN evaluation is expensive
- Accuracy alone misses hardware cost
- LLM's proposals may violate search constraints

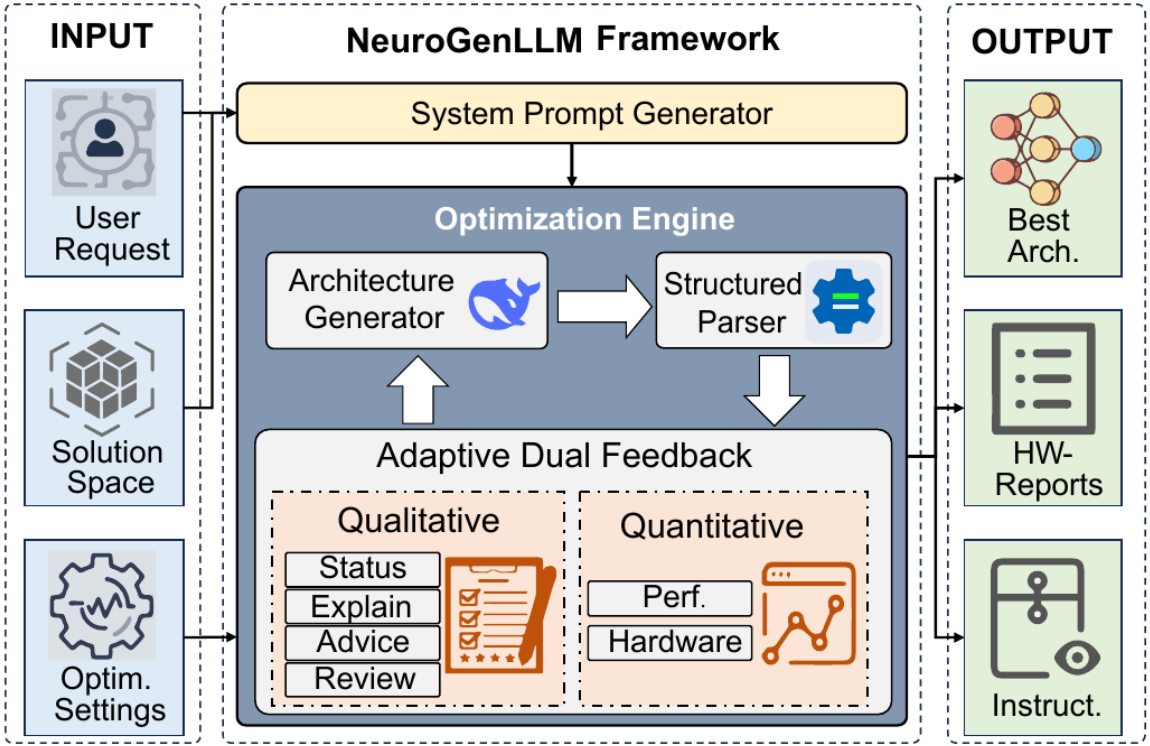
NeuroGenLLM

- Structured generation with verification
- Feedback from accuracy and hardware metrics
- Search history guides exploration

more sample-efficient

Goal: competitive accuracy with lower search and hardware cost

NeuroGenLLM framework



Generate, verify, evaluate, and refine

What adaptive dual feedback means

Two forms of evidence guide every new proposal

QUANTITATIVE

Which candidate is better?

Evidence
Accuracy and hardware proxies

Result
Evaluation logs for ranking candidates

+

QUALITATIVE

What should change next?

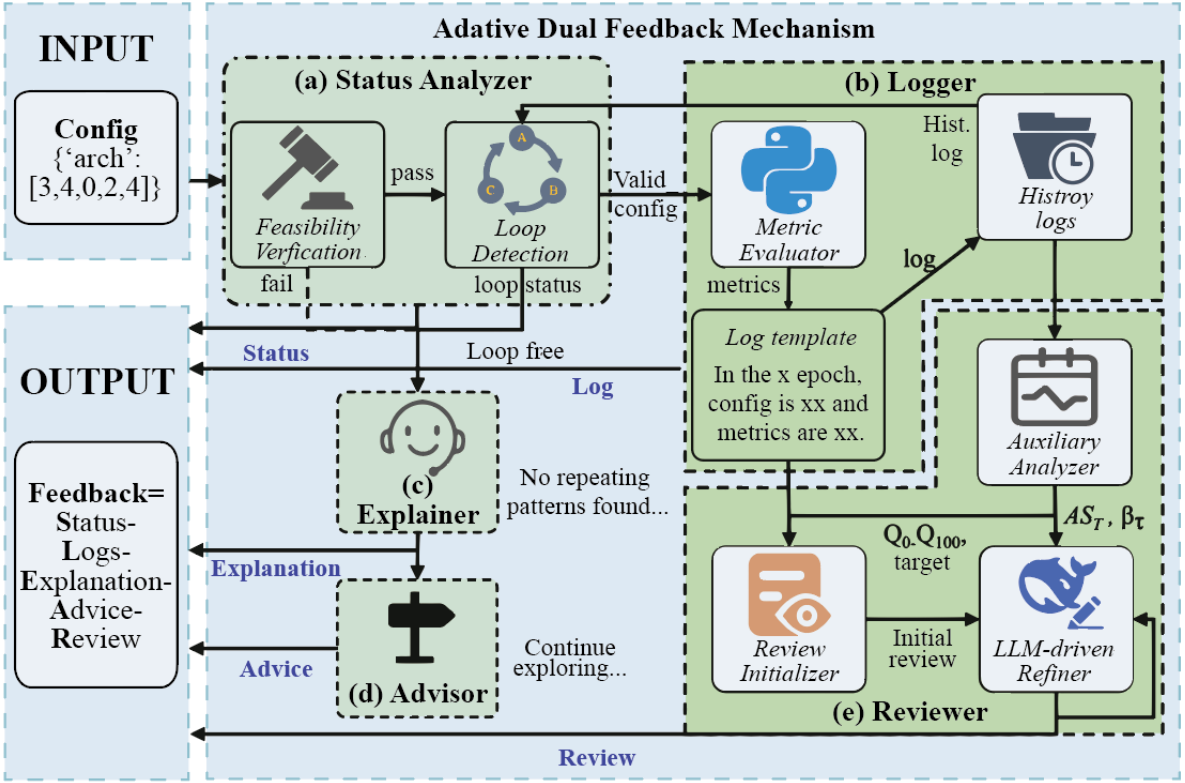
Evidence
Constraint check, search history, and repeated patterns

Result
Explanation, advice, and review

NEXT ARCHITECTURE PROPOSAL

Numbers rank candidates. Reasoning steers the search.

Adaptive Dual Feedback (ADF) architecture



Five modules

1. Status Analyzer
2. Logger
3. Explainer
4. Advisor
5. Reviewer

Hardware-aware feedback

Composite objective

$$F(A) = Acc \times [(Spikes/Spikes_0) \times (SynOps/SynOps_0) \times (Params/Params_0)]^{\lambda}$$

Effect of λ on CIFAR10

λ	Acc. (%)	Spikes	SynOps	Params	Energy	Latency
0	93.14	376K	407M	5.13M	9.61	22.24
-0.1	92.98	312K	312M	3.12M	7.39	19.24

λ = -0.1 keeps the accuracy loss to 0.16% while reducing every hardware proxy

Experimental setup

Datasets

Static image

- **CIFAR10**
- **CIFAR100**
- **Tiny-ImageNet-200**

Neuromorphic

- **CIFAR10-DVS**

Protocol

- 20 architecture evaluations
- DeepSeek-V3, temperature 0
- One NVIDIA Quadro GP100

Main comparisons

- Random
- AutoSNN
- GENIUS

Accuracy, SynOps, and search time

Accuracy and hardware results

AutoSNN comparison under the reported settings

Dataset	Acc. Auto	Acc. Ours	Δ pp	SynOps Auto	SynOps Ours	Reduction
CIFAR10	92.54	92.98	+0.44	474M	312M	34.2%
CIFAR100	69.16	70.43	+1.27	716M	536M	25.1%
Tiny-ImageNet	46.79	47.13	+0.34	756M	610M	19.3%
CIFAR10-DVS	72.50	73.03	+0.53	205M	199M	2.9%

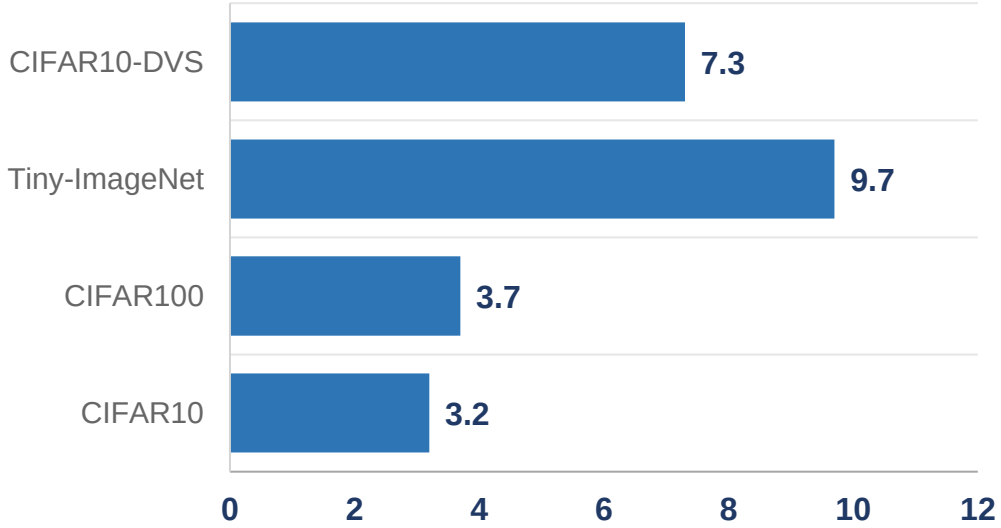
20
architectures evaluated
AutoSNN evaluates 400

+0.34 to +1.27 pp
accuracy gain
across all four datasets

2.9%-34.2%
SynOps reduction
relative to AutoSNN

Search efficiency and quality

Search-stage speedup over AutoSNN



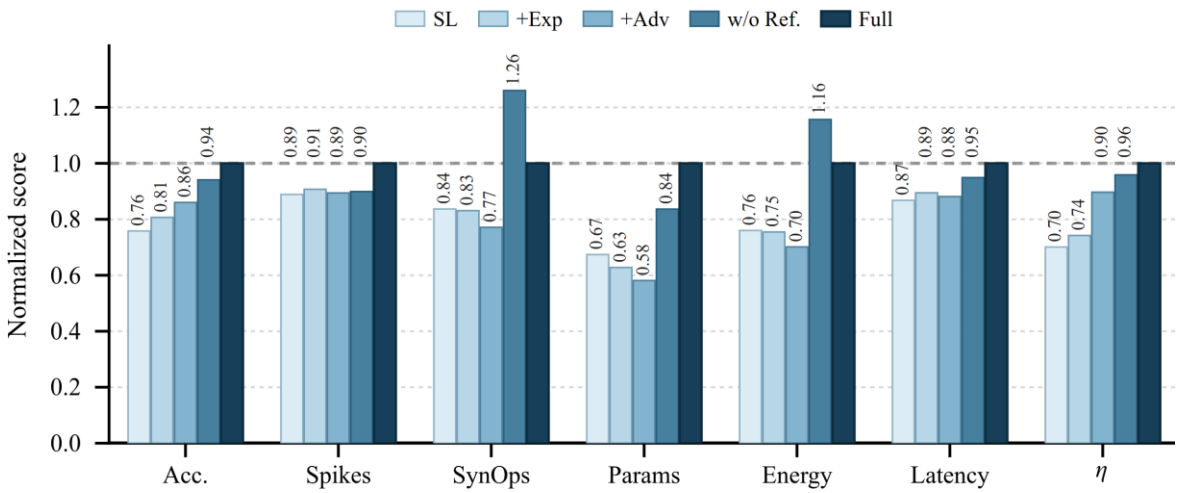
Compared with GENIUS

Metric	GENIUS	Ours
Exploration rate	80%	97%
Valid generation	85%	95%
API cost	$0.06	$0.10

NeuroGenLLM achieves higher accuracy on all four datasets

Ablation study

Adaptive Dual Feedback (ADF)



Quantitative and qualitative feedback provide complementary gains

Full ADF versus logging (SL)

24.2%

lower error rate

>24% fewer SynOps

>32% fewer parameters

Exploration rate: 0.97

Conclusion

makes SNN architecture search more sample efficient

20 evaluations	A small search budget
Adaptive feedback	Accuracy, hardware, and search history
3.2x-9.7x faster	Higher accuracy and fewer SynOps than AutoSNN

Limitation: energy and latency use hardware proxy models

Thank you - Q&A

Questions?

NeuroGenLLM

A Fast Spiking Neural Architecture Generation Framework
via LLM-Guided Search

IEEE SMC 2026 | BMI Workshop | Paper 256

References

- [1] Roy et al., Towards spike-based machine intelligence with neuromorphic computing, Nature, 2019.
- [3] Kim et al., Neural architecture search for spiking neural networks, ECCV, 2022.
- [4] Na et al., AutoSNN: Towards energy-efficient spiking neural networks, ICML, 2022.
- [5] Yan et al., Efficient spiking neural network design via neural architecture search, Neural Networks, 2024.
- [8] Sun et al., SpikeNAS-Bench, IEEE Transactions on Artificial Intelligence, 2025.
- [9] Davies et al., Loihi: A neuromorphic manycore processor with on-chip learning, IEEE Micro, 2018.
- [10] Nasir et al., LLMatic: Neural architecture search via large language models, GECCO, 2024.
- [12] Zheng et al., Can GPT-4 perform neural architecture search?, arXiv:2304.10970, 2023.
- [13] Zhang et al., MLCopilot, EACL, 2024.
- [15] Ji et al., Survey of hallucination in natural language generation, ACM Computing Surveys, 2023.