



Motivation

Partners want open-source SLMs on embedded MCU/NPUs without retraining or QAT from scratch.

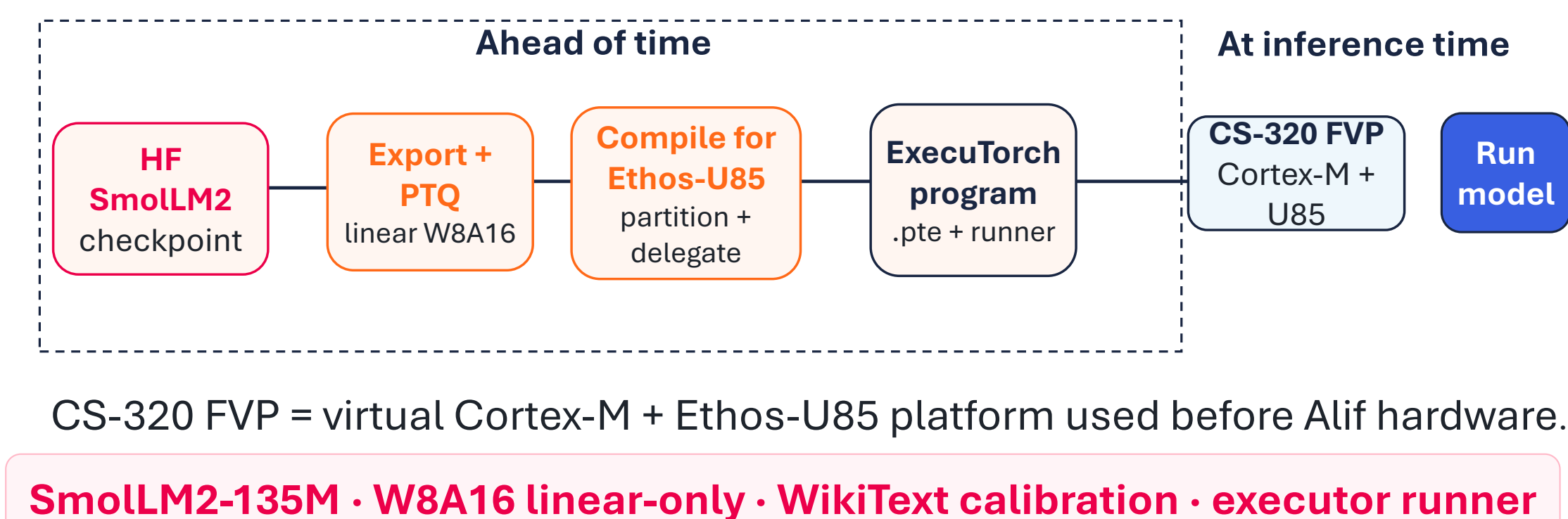
- Arm developed and evaluated a selective, backend-aware W8A16 PTQ approach
- FP32/BF16 keeps quality but exceeds small-device memory budgets
- Full-graph W8A8 is smaller, but generation collapses
- Goal: reduce size while keeping usable output

Target model
SmolLM2-135M

Target NPU
Ethos-U85

Deployment
Alif-class

End-to-End ExecuTorch Workflow



Method: Why Linear-Only W8A16?

Arm approach: quantize only torch.nn.Linear layers

- Linear layers dominate transformer compute and memory
- W8A16 maps cleanly to the Ethos-U quantized path
- Avoids quantizing RoPE, indexing, softmax and attention glue
- Calibration uses representative WikiText activations; no retraining

Accuracy baseline: smaller is not automatically better — full-graph W8A8 gives unusable text; linear-only W8A8 roughly doubles perplexity. Selective W8A16 is the practical baseline.

Key Result: 63% Smaller + Usable Generation

Arm W8A16: ~63% smaller, <+3.1 PPL, usable output

623 → 228 MB

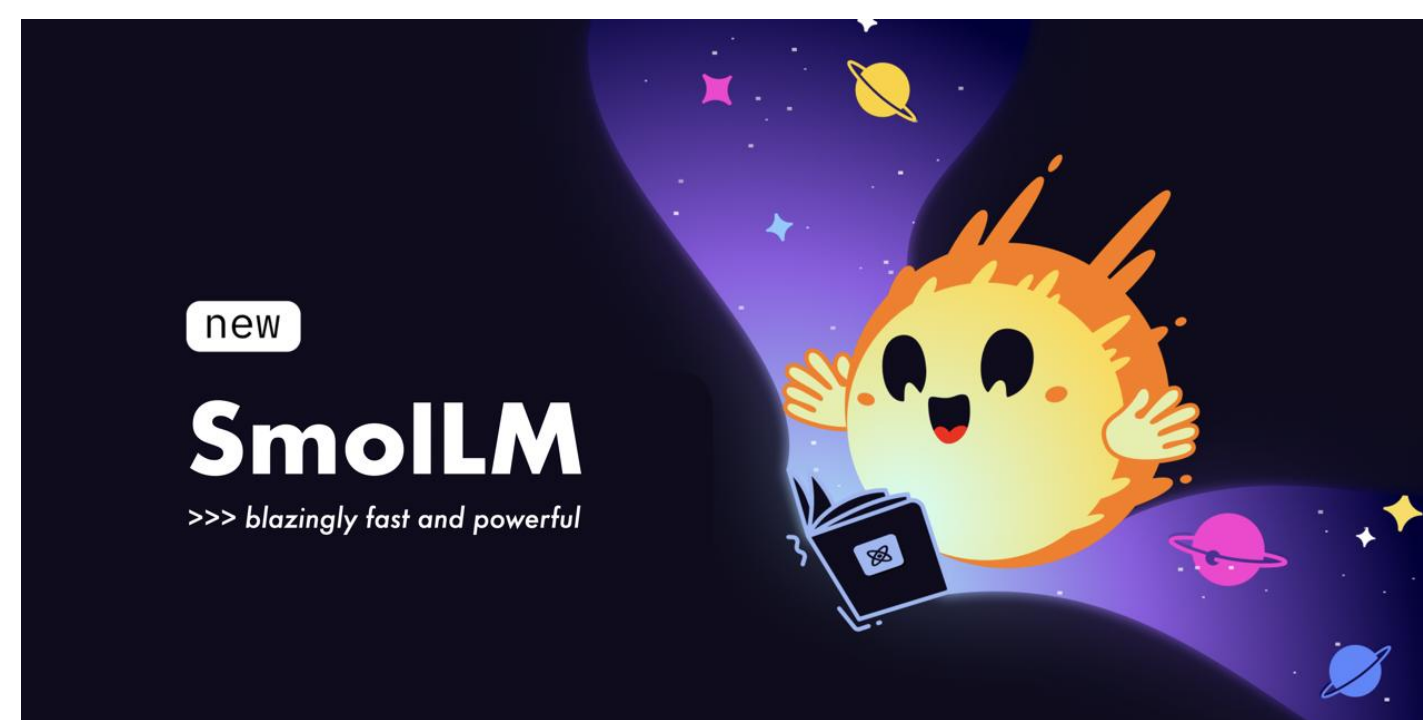
model artifact size
W8A16 linear-only

42.5 → 45.0

WikiText PPL @ ctx64

~63% smaller

usable output
full W8A8 failed



Qualitative Generation: FP32 vs W8A16

Greedy decoding shows W8A16 keeps short generations coherent instead of collapsing into rubbish output.

Prompt: Once upon a time in a small village...

FP32: Sam and Alex find an old map that leads to a hidden treasure chest.

W8A16: Sam and Alex explore together and find a strange box with buttons.

Prompt: What's the capital of France?

FP32: The capital is Paris. Then continues with a generic Q&A pattern.

W8A16: The capital of France is Paris. It continues with plausible city details.

Why This Matters for Arm Edge Devices

1.7 tok/s

TinyStories-42M on Alif E8
CortexM55+EthosU85

2.5 tok/s

TinyStories-42M on Alif E8
CortexA32+EthosU85

125 MB

flash budget reference

SmolLM2-135M is larger than today's smallest MCU demos, but the deployment pressure is the same: every MB saved matters.

Selective W8A16 helps move from toy text models toward more capable open-source SLMs on resource-constrained Arm edge devices.

Arm evaluation target: Alif E8 dev kits

Developer Takeaway

How to preserve quality:

1. Start from a Hugging Face checkpoint
2. Export with ExecuTorch
3. Calibrate with representative text
4. Quantize backend-compatible Linear layers
5. Compare PPL and generated text on target/FVP

Lesson: smaller is not always better. Reduce model size while preserving usable output on the target hardware.

Conclusion & Next Steps

Arm's selective W8A16 approach cuts SmolLM2-135M artifact size by ~63% while preserving usable generation where full W8A8 fails.

Future work

- Continue KV-cache memory and mutation optimization
- Multi-token prefill for better prompt throughput
- SpinQuant/Hadamard-style rotations for lower bits
- Push open-source SLMs toward embedded NPUs

Goal: make practical SLM deployment reproducible for resource-constrained PyTorch users.