

01 Why dynamic W8A8?

Motivation

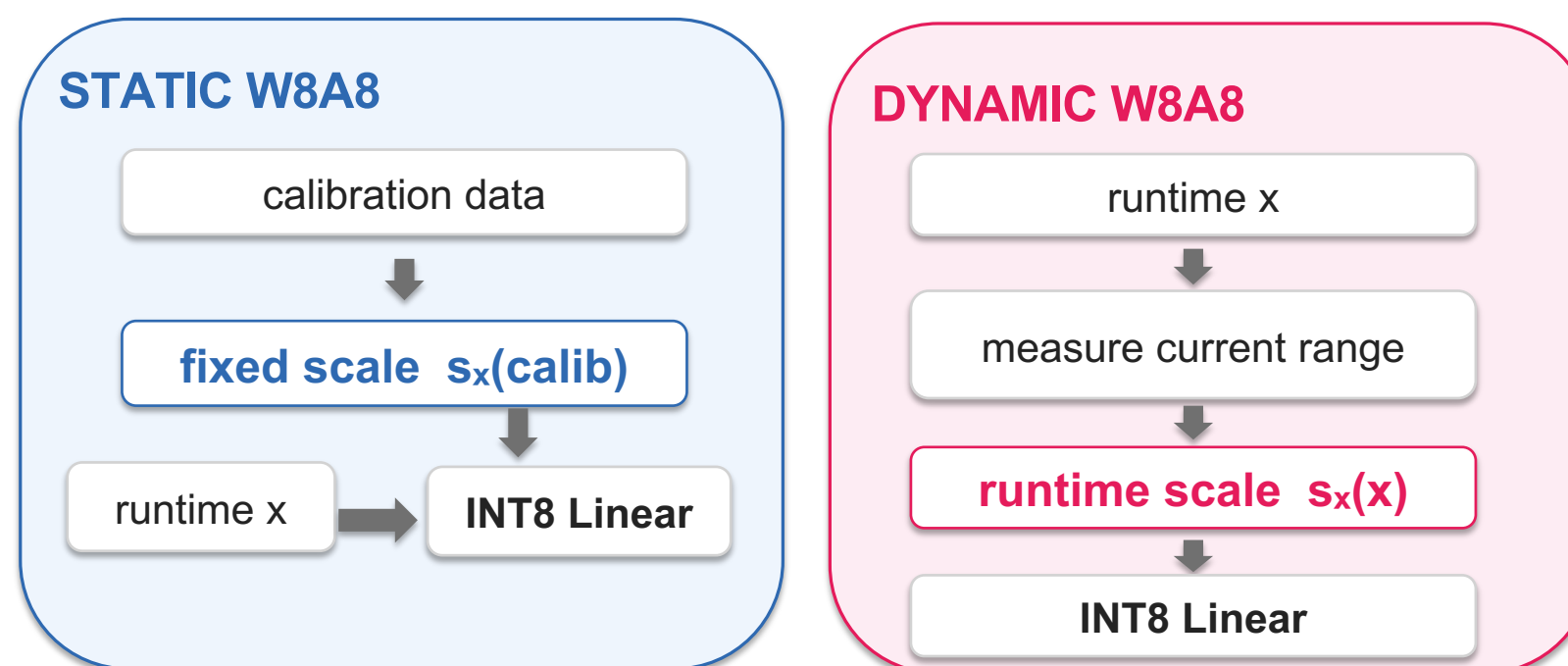
Static W8A8: activation scale is fixed from calibration data.

Range shift: runtime values can saturate or waste INT8 precision.

Dynamic W8A8: recompute the activation scale from the current tensor; keep weights static INT8.

Goal: keep INT8 Linear execution without a fixed activation calibration range.

02 Static vs dynamic



Tradeoff: small runtime overhead for input-adaptive activation scaling.

03 Dynamic W8A8 math

$$s_x = \max\left(\frac{\max_i |x_i|}{127}, \epsilon\right)$$

$$q_x = \text{clip}(\text{round}(\frac{x}{s_x}), -127, 127)$$

$$A = q_x q_w^T \text{ [INT32]}$$

$$y = \text{FP32}(A) (s_x s_w) + b$$

s_x runtime • s_w static • q_x, q_w INT8 • A - INT32 accumulation

Why Linear

Linear layers dominate compute in transformers and MLPs and map directly to efficient INT8 matrix multiplication.

NOVELTY

Runtime qparams survive as values

PT2E dynamic QDQ → Arm/TOSA INT8 MATMUL → VGF execution

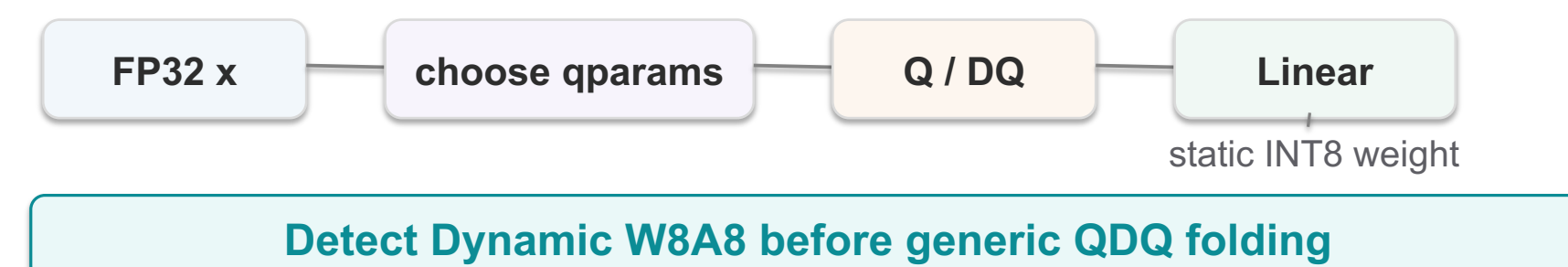
Preserve s_x through generic QDQ folding; reuse it for runtime FP32 rescale + bias.



04 How it works

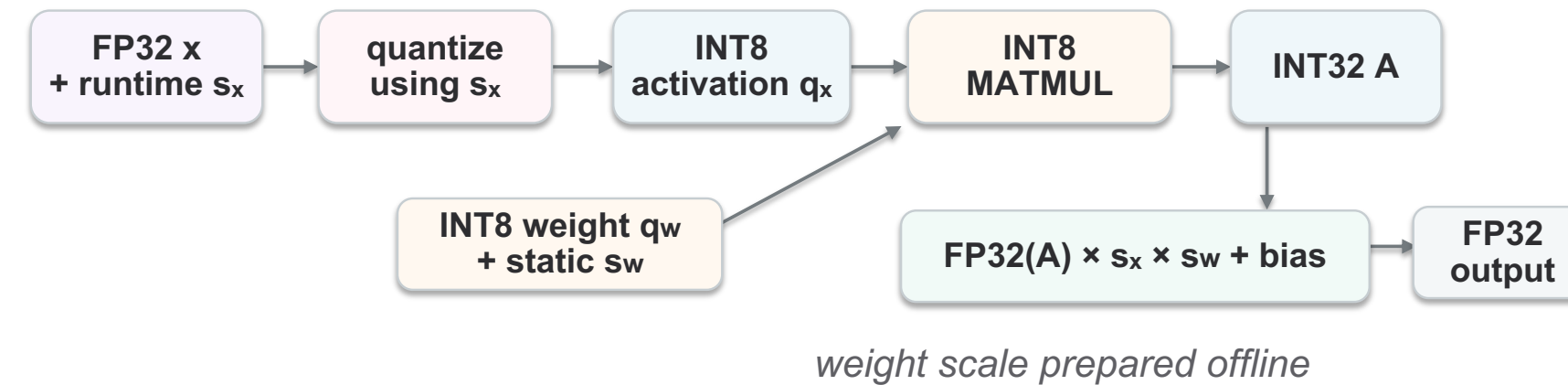
PT2E dynamic QDQ

Runtime qparams depend on x; keep s_x live until backend-specific lowering.



Arm/TOSA lowering

Replace dynamic QDQ + Linear with INT8 MATMUL, INT32 accumulation, and FP32 rescale + bias.



05 Compiler challenges

1 Preserve runtime values

Static-qparam passes must not fold or disconnect runtime s_x .

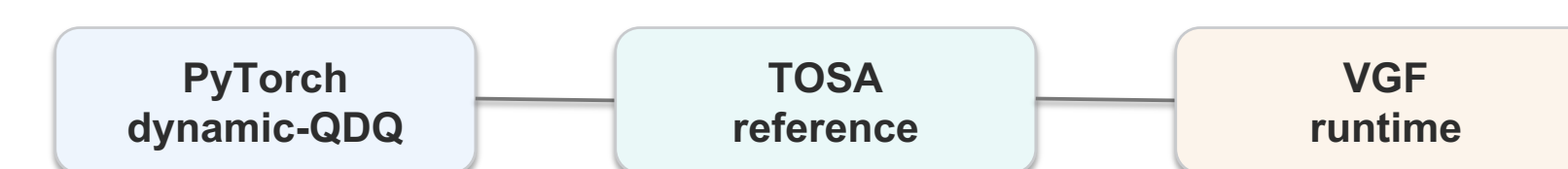
2 Match real PT2E graphs

Handle explicit QDQ and constant-folded INT8 weights.

3 Bias depends on runtime scale

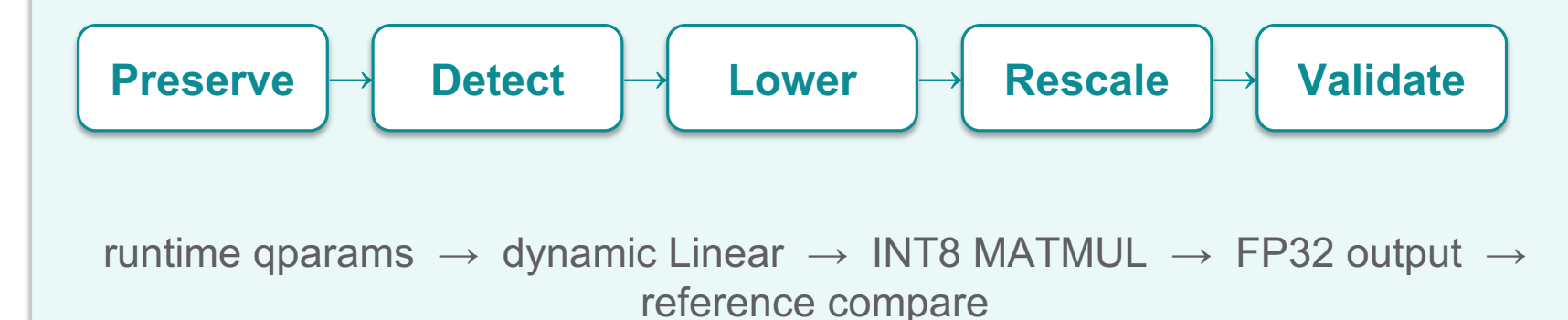
Because s_x is runtime, add bias after FP32 reconstruction.

Validation



Structural tests verify dynamic qparams and INT8 TOSA MATMUL; numerical tests compare TOSA/VGF execution with the PT2E dynamic-QDQ reference.

Reusable backend pattern



06 Results

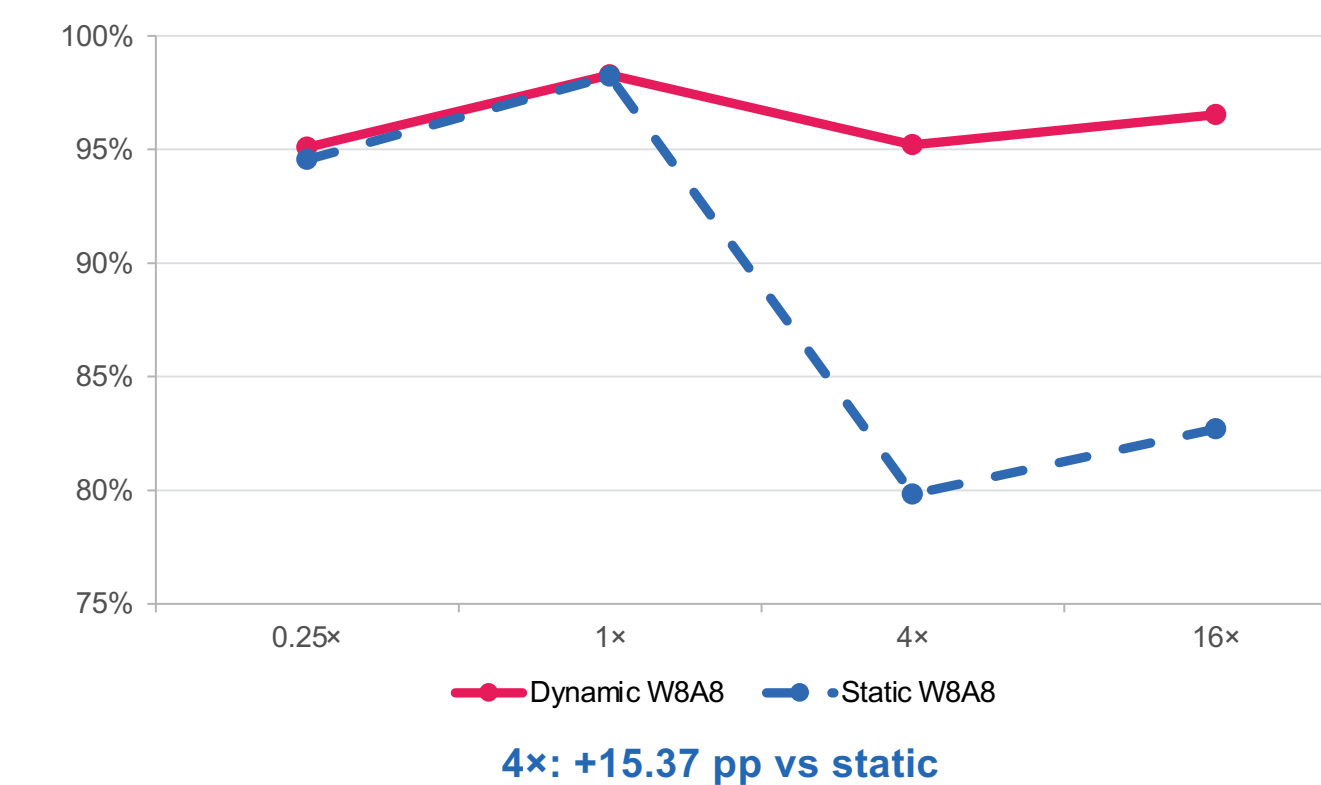
DeiT-Tiny • full test split • n = 3,669

+15.4 pp FP32 agreement at 4×

95.2% Dynamic vs **79.8% Static**

Fidelity ≠ accuracy — both are shown below

Agreement with FP32 predictions



Top-1 accuracy

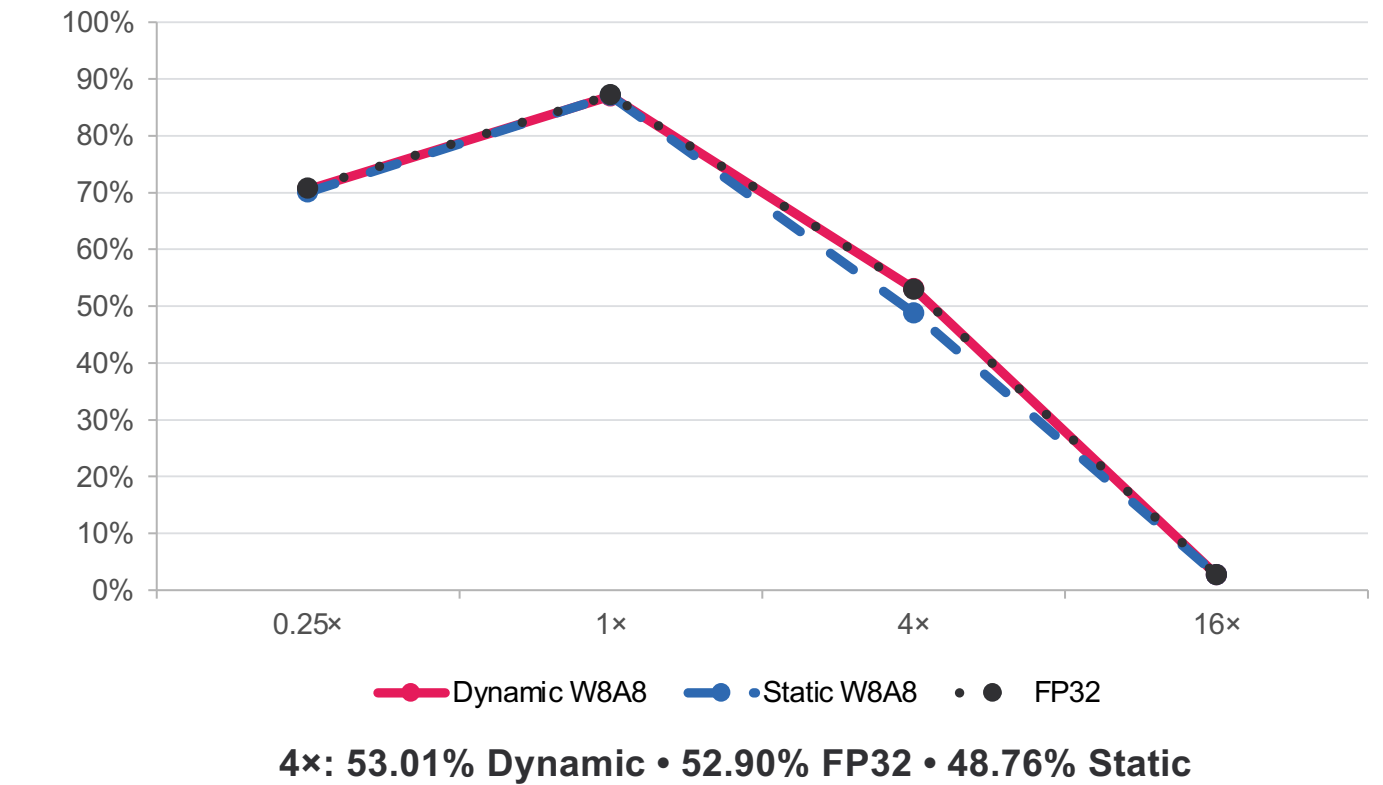
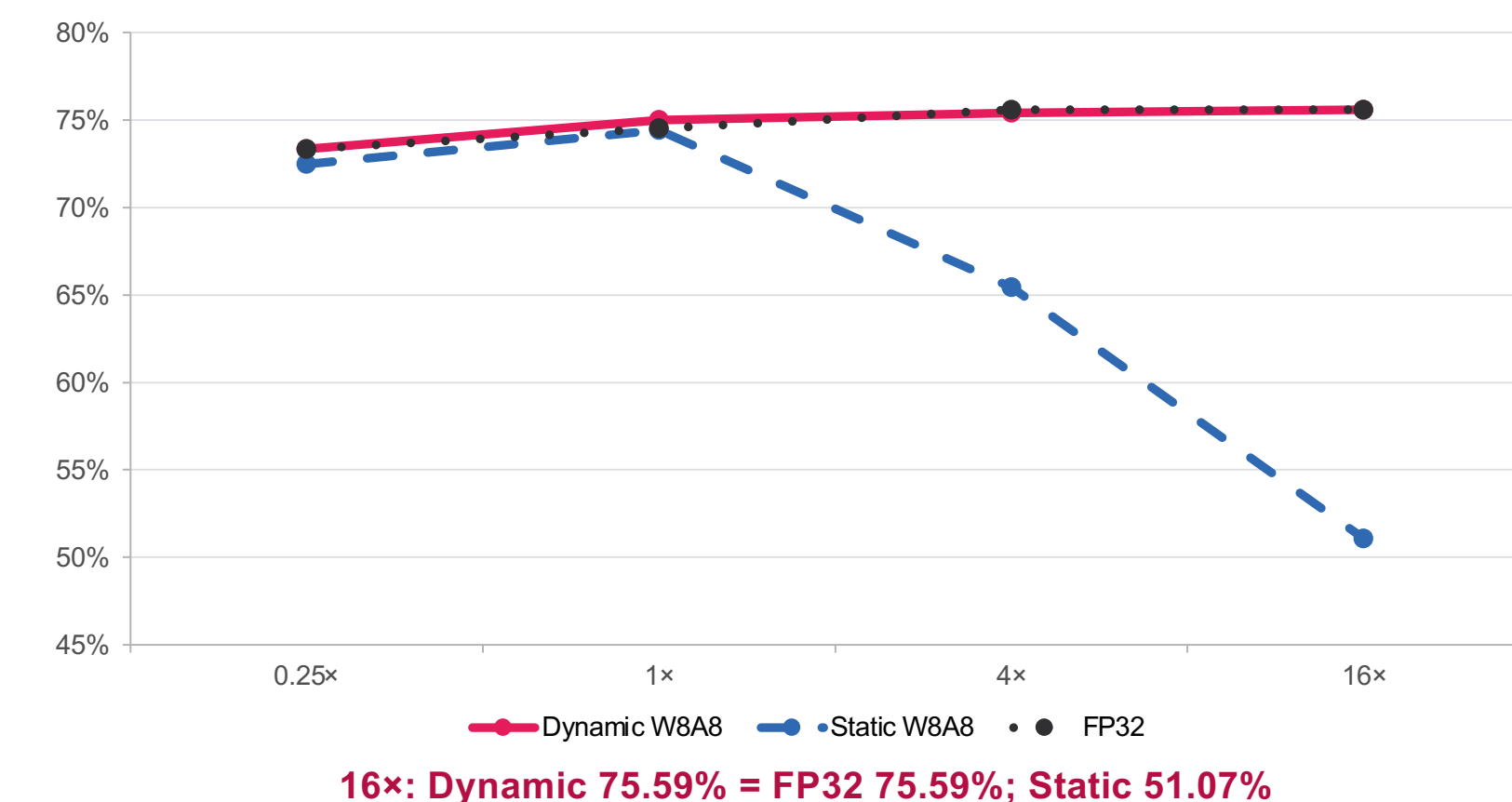


Image MLP — top-1 accuracy



3.32× smaller PTE

21.60 → 6.51 MiB

72 / 72 rank-3 Linear

all transformer Linear/AddMM exercised

1 delegated region

no Linear / choose_qparams / QDQ remain

18.2× lower MAE

synthetic DLRM @16× vs static

07 Takeaway

Runtime activation scales. Static INT8 weights. Integer MATMUL.

Dynamic W8A8 stays closer to the same-scale FP32 reference when activation ranges shift.

It reduces added quantization error; it does not make the FP32 model itself invariant to severe input rescaling.

Backend pattern: preserve → detect → lower → rescale → validate

TRY IT



Dynamic W8A8 tutorial

ExecuTorch Arm VGF

Scan for code, export steps, and VGF execution.