



A Generalized Optimization Engine (GOE) For Edge AI Inference Acceleration



GOE-compressed LLMs run on a GPU-less edge CPU: proper 8-bit quantization is **~1.9× smaller** and **~1.7× faster** with accuracy essentially unchanged, while a naive 8-bit scheme collapses to chance. The compression **method**, not just the bit-width, decides what survives deployment.

1. PROBLEM

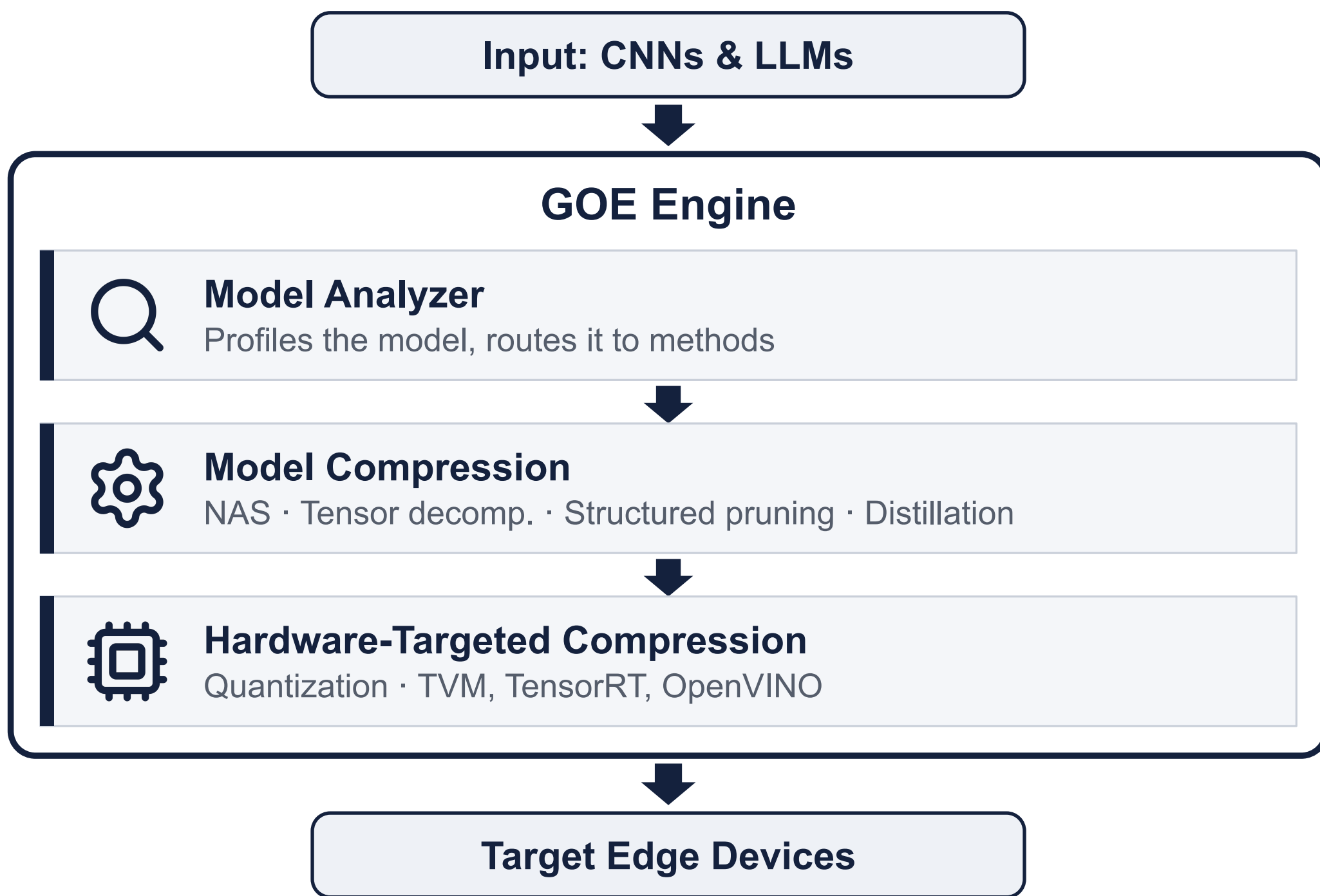
- Tactical edge is **dynamic, heterogeneous, resource-constrained**; LLMs are hard to deploy there.
- Existing compression methods work in a **narrow, model- or platform-specific** context, often with accuracy decay.
- Goal: one **model- and hardware-agnostic** engine.

2. FORMULATION



Maximize **Performance**(model, quantization, pruning, hyperparameters) subject to budgets on **inference time, memory, and energy** on the target device. A multi-objective variant also maximizes compression rate.

3. GOE APPROACH



Split by model scale: supernet NAS for CNNs; compression routes for LLMs.

4. CNN RESULTS



>25×

ResNet 50 throughput with TorchAO INT8, accuracy unchanged (81.0%)



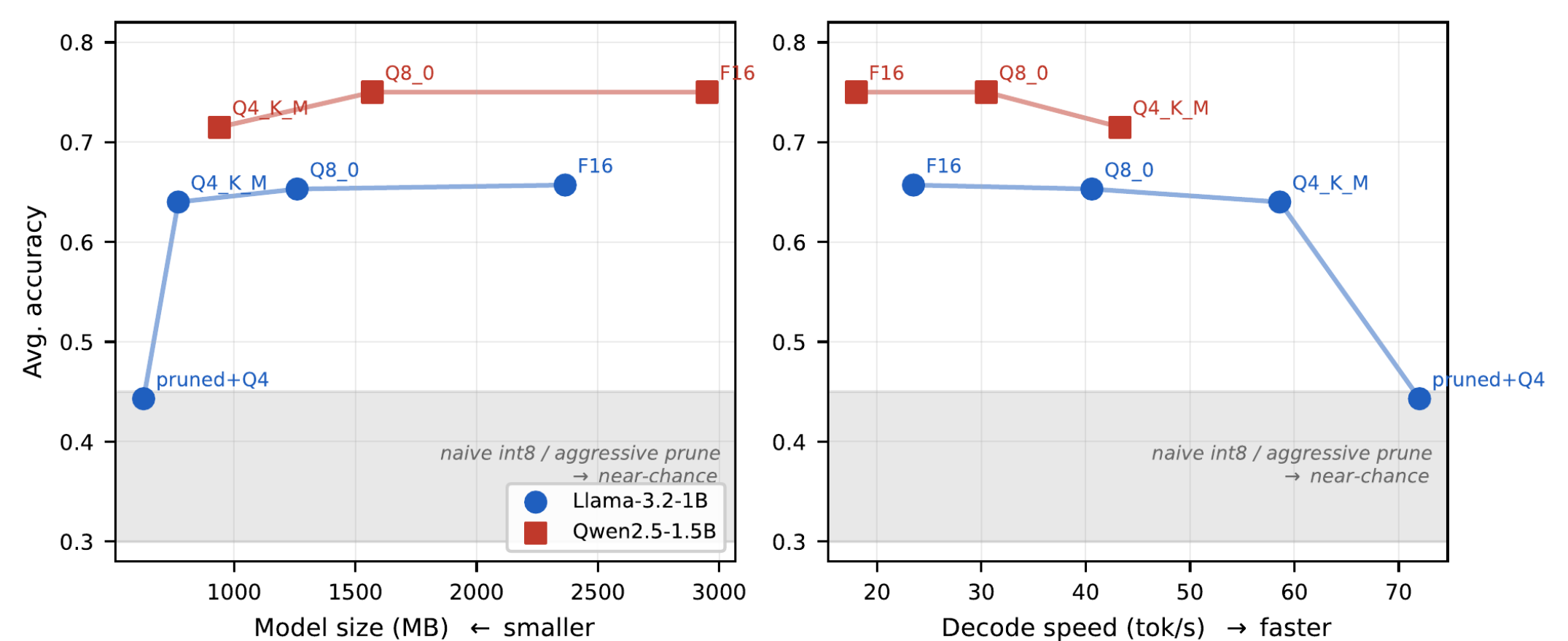
97.7→6.2

ResNet 50 size (Mb), 75% pruned, 5 epochs; accuracy matched or beat base

- NAS on ResNet 50: top-10 subnets all **≥2× compressed**; quantization trims ~10% more.
- Caveat: MobileNet v3 Small, already compact, drops 68 → 38% under INT8.

5. EDGE-CPU DEPLOYMENT OF LLMs

- Llama-3.2-1B, Qwen2.5-1.5B on Intel Core Ultra 7 265U, no GPU (llama.cpp / GGUF).
- Routes: quantization (Q8_0, Q4_K_M), structured pruning + recovery.



Avg. accuracy vs. size and speed. Quantization holds accuracy near F16; pruning and naive int8 fall into the near-chance band (shaded).

6. KEY RESULTS

1.9× / 1.7×

Q8_0: smaller / faster, accuracy essentially unchanged

~3× / 2.5×

Q4_K_M: smaller / faster, for a few accuracy points

Llama-3.2-1B on the edge CPU

Variant	MB	tok/s	ARC-e	PIQA	HS	CSQA
F16	2365	23.5	.620	.760	.627	.620
Q8_0	1260	40.6	.613	.753	.627	.620
Q4_K_M	770	58.6	.613	.760	.600	.587
pruned+Q4	627	72.0	.440	.640	.500	.193

Qwen2.5-1.5B on the edge CPU

Variant	MB	tok/s	ARC-e	PIQA	HS	CSQA
F16	2950	18.0	.747	.780	.667	.807
Q8_0	1570	30.5	.753	.773	.667	.807
Q4_K_M	940	43.3	.747	.747	.633	.733

ARC-e = ARC-Easy, HS = HellaSwag, CSQA = CommonsenseQA.

7. TAKEAWAYS

- Method beats bit-width:** naive dynamic 8-bit, same width as Q8_0, collapses to chance through activation-outlier error.
- Aggressive pruning** buys speed but nears chance even after recovery.
- GOE's job:** route each model to a method that deploys *and* works.
- Next:** multimodal inputs and on-the-fly reconfiguration mid-mission.

FOR FURTHER INFORMATION:

U.S. ARMY COMBAT CAPABILITIES
DEVELOPMENT COMMAND
ARMY RESEARCH LABORATORY:

DEVCOM.ARMY.MIL

POINT OF CONTACT:

*Venkat R Dasari, Venkateswara.r.Dasari.civ@army.mil
Team: *Vinod Mishra, *Jakob Adams and **Brian Jalaian
* DEVCOM Army Research Laboratory and **University of West Florida

Distribution Statement A: Approved for public release; distribution unlimited

