TensorRT Roadmap

	Oct'25	Jan'26	Q1'26	
TRT Version	10.12 -10.14	10.15	10.16	11.0
Performance	- NVFP4 gemm perf optimization - Blackwell/Thor perf optimization	- Continuous Blackwell/Thor perf optimization	- Selected auto model perf optimization based on customer requests - Recycs model perf optimization - Audio2Text model perf optimization	
Ease of Use	- Improve debuggability: INetwork API Capture & Replay - Improve debuggability: Share internal tensor value dump knob with users	- Provide best Practice workflow for Strongly Typing -Improve engine graph visibility: add query function for static weights size - Plugin and QDP enhancement	- Provide IpluginV2 ->> IpluginV3 migration guidance	-[Frontend] provide Torch-TRT as the production ready frontend and officially promote it in 11.0 -[Frontend] [Github] provides step by step guidance for different workflows -[Frontend] Promote AItune for direct HF model importing -[Debuggability] Accuracy Awareness API
Functional	- MHA API	- QDQ placement autotune - Add RoPE API		- Enable Multi GPU on Datacenter - MoE API - KVcache API
	Work with ModelOpt on QDQ node placements: bugs, placement heuristics, perf-tuning			