

TensorRT Roadmap

	May, 2026	June, 2026	July, 2026	Q3, 2026
<i>TRT Version</i>	11.0	11.1	11.2	11.x
<i>Performance</i>	- Selected auto model perf optimization based on customer requests	- Perf optimization per customer request	- Perf optimization per customer request	- Perf optimization per customer request
<i>Ease of Use</i>	- [Frontend] provide Torch-TRT as the frontend in 11.0 for in-framework customers - [Frontend] [Github] provides step by step guidance for different workflows - [Frontend] Promote Altune for direct HF model importing	- [Debuggability] Performance Awareness API - Update QDP sample to demonstrate autotuning - Create new sample for QDP + CuteDSL	- OSS TensorRT Model Connect	- Provide checkpoint visibility with detailed fusion optimization info and tracking metadata
<i>Functional</i>	- QDQ placement autotune - Enable Multi GPU on Datacenter	- Support Python 3.14 - Support Ubuntu 26.04	- Enable DLA Strongly typing	- Cross compilation + remote autotune (expand scope from Auto safety to TRT standard) - Native support FFT/DFT - Add GridSample3D as native op - Multi GPU enabled on embedded platform - Hardware compatibility in ARM - Support FP8 fallback support for older GPUs
<i>SKILL/Agentic</i>		- SKILLS for best perf optimization - SKILLS for best perf visualize - SKILLS for weakly to strongly migration guidance	- SKILLS for 3D medical imaging best practices	