PEFT 0.22: PaCA fine-tuning, new LoRA initialisations and breaking changes

27 November 2026 · 5 min read · blogs.iictp.org/peft-0-22-paca-lora-upgrade-guide

Short answer: PEFT 0.22 is a release of Hugging Face's parameter-efficient fine-tuning library, published on 19 November 2026, that adds the PaCA method, Astra and NoRA initialisation for LoRA and wider parallel training support. It also changes several defaults and requires Python 3.11 or newer. It matters because some existing training and loading scripts will warn, fail or behave differently after upgrading.

TL;DR The short version

- PEFT 0.22, released on 19 November 2026, adds the PaCA method, Astra and NoRA initialisation for LoRA, and wider parallel training support.
- It changes several defaults and requires Python 3.11 or newer.
- Some training and loading scripts will warn, fail or behave differently, so pin versions and test before upgrading.

Read online, listen to it, or change the text size: https://blogs.iictp.org/peft-0-22-paca-lora-upgrade-guide

What is PEFT 0.22?

Parameter-efficient fine-tuning adapts a large model by training a small number of parameters. LoRA is the best-known method: it adds small low-rank matrices beside frozen weights. This release adds a new method, new ways to start LoRA weights and many engineering changes. The maintainers say the next release will be version 1.0.

How the new methods work

- **PaCA** trains randomly chosen input columns of the pretrained weights instead of adding adapter layers, and stores only partial activations for the backward pass.
- **Astra initialisation** uses a small calibration set to start LoRA in low-variance directions of each layer's output. Set init_lora_weights to astra and run the preprocessing step first.
- **NoRA initialisation** normalises columns of the random LoRA A matrix; set lora_alpha equal to the rank.

What is new for distributed training and operations

- LoRA supports a column-wise embedding plan for tensor parallelism, and replicated weights start synchronised.
- FSDP2 and HSDP training, export and merging received fixes; the notes advise Transformers 5.17 or newer.
- A health check method summarises adapter state, data types and trainable parameters.

- A helper builds a placeholder adapter so that differently shaped LoRA adapters can be hot-swapped.
- More adapter types can be converted to LoRA.

Results reported by the source

- Citing the PaCA paper, the notes give 22% less training time, 16% less memory and 23% longer sequences than LoRA, with comparable accuracy.
- They say PaCA helps most when training time or activation memory is the bottleneck, helps less with few tokens per step or light target modules, and supports only linear layers.
- The NoRA description reports faster and better convergence than standard LoRA initialisation in experiments.

Breaking changes and open questions

- Loading an adapter now warns about missing and unexpected keys by default; strict_adapter_check controls whether it errors or ignores them.
- modules_to_save now matches whole sub-module names, not any suffix.
- k-bit preparation now uses the Transformers default for activation checkpointing.
- low_cpu_mem_usage keeps data types.
- A fix to OFT convolution layers can change outputs of existing checkpoints used without merging.
- The quoted gains come from the method’s paper, so verify them on your model and data.

Sources were read on 27 November 2026. Products, rules and data change, so check the linked originals before you rely on any detail.

How to upgrade

1. Move the environment to Python 3.11 or newer.
2. Freeze current versions and save evaluation outputs.
3. Search configurations for modules_to_save and OFT.
4. Load every production adapter with strict checking set to error and fix mismatches.
5. Run one short training job and compare loss, memory and time.
6. Try PaCA or a new initialisation as a separate experiment against your LoRA baseline.

What it means for fine-tuning teams

The stricter load warnings can expose adapters that were silently half-applied, which is worth finding. Change one thing at a time so a difference can be traced to the library, the initialisation or the data. For library changes in a sister project, see TRL 1.19.

Frequently asked questions

- **What is new in PEFT 0.22?** PaCA, Astra and NoRA initialisation, parallel training improvements and a health check.
- **Does PEFT 0.22 support Python 3.10?** No. It requires Python 3.11 or newer.

- **Is PaCA better than LoRA?** The paper cited reports efficiency gains with comparable accuracy, in the cases it describes.
- **Will old checkpoints change?** Some OFT convolution checkpoints may produce different unmerged outputs.

Related reading

- TRL 1.19 distillation and packing
- Predicting fine-tuning time
- RAG vs fine-tuning
- Local LLM Practitioner certification

> Related credential: Local LLM Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Hugging Face. PEFT release v0.22.0. 19 November 2026. github.com/huggingface/peft/releases/tag/v0.22.0

Published by IICTP, an independent technical credential body, on 27 November 2026. The current version of this post is at https://blogs.iictp.org/peft-0-22-paca-lora-upgrade-guide. More guides: blogs.iictp.org.