# Чек-лист сжатия для edge — дистилляция + квантизация



## 1 - Два рычага, делающих модель меньше

```
Квантизация  оставить модель, хранить числа меньшим числом бит (FP32/16 -> INT8/INT4). Дёшево.
Дистилляция  обучить НОВУЮ меньшую модель-'ученика' копировать большую 'учителя'. Дорого создавать.
Используйте ОБЕ: сначала дистилляция, затем квантизация ученика -> экономия объёма + скорости множится.
```

## 2 - Арифметика памяти (модель на 7 млрд параметров)

```
FP32 (4 Б/вес)=28 ГБ   FP16 (2 Б/вес)=14 ГБ   INT8 (1 Б/вес)=7 ГБ   INT4 (0.5 Б/вес)=3,5 ГБ
Потолок 8 ГБ edge (Jetson Orin Nano): FP16 не влезет; INT8 едва влезает; INT4 с запасом.
INT8 ~= в 4 раза меньше FP32, обычно <1% точности. INT4 нужен AWQ/GGUF, чтобы остаться точным.
```

## 3 - Методы квантизации (узнавайте в спецификации подрядчика)

```
INT8 + калибровка   8-бит; edge через TensorRT; <1% потери. Безопасный дефолт для CV-моделей.
AWQ                 4-бит; защищает ~1% значимых каналов; качество близко к FP16; для VLM/LLM.
GPTQ                4-бит; послойная коррекция ошибки; второй зрелый 4-битный путь.
GGUF Q4_K_M         ~4.5-бит; llama.cpp на ноутбуке/Mac/коробке; ~2% качества за ~40% памяти.
FP8 / NVFP4         8-/4-бит float на NVIDIA; NVFP4 на Blackwell: <=1% потери, до 4x скорости FP8.
PTQ vs QAT          PTQ сначала (минуты, без дообуч.). QAT если потеря PTQ слишком велика.
```

## 4 - Видео-функция это конвейер - сжимайте каждый этап

```
Детекция/сегментация (YOLO, SAM)  INT8 через TensorRT - PTQ сначала, QAT если точность падает.
Транскрипция (речь->текст)        дистиллир. модель (Distil-Whisper), затем квантизация до INT8.
Описание/резюме (VLM/LLM)         4-бит веса через AWQ или GGUF - крупнейшие модели, главный выигрыш.
```

## 5 - Перед сжатием спросите:

- [ ] Нужна ли эта функция на edge вообще (объём, полоса, задержка, приватность) - или облако подойдёт?
- [ ] Попробовали ли сначала дешёвый рычаг: post-training INT8 квантизацию, до дистилляции или QAT?
- [ ] Большие языковые / vision-language этапы на 4 битах (AWQ или GGUF), где память кусается сильнее?
- [ ] Для речи берём дистиллированную модель (Distil-Whisper), а не полного учителя - затем квантизуем?
- [ ] Перемеряли ли точность на СВОИХ кадрах после каждого шага, а не доверяли среднему по бенчмарку?
- [ ] Проверили ли редкие-но-важные классы (оружие, падение, дефект), а не только среднюю точность?
- [ ] Влезает ли модель в потолок памяти устройства (часто 8 ГБ) с местом под рабочую память?
- [ ] Экспортирована ли модель в правильный рантайм (TensorRT/INT8 на Jetson; Core ML / NNAPI на телефонах)?