Which GPU, spot or on-demand: 12 AI workloads on one page

Match the card to the memory footprint, the market to the interruption tolerance - not the workload's reputation

WORKLOAD	OPTIMAL GPUS	PLACEMENT	THE CONSTRAINT THAT DECIDES
Support bots	L4 / A10 · L40S / A100 for heavier models or high concurrency	**ON-DEMAND — users are waiting on the response**	Concurrency: every open chat holds its own KV cache
Internal knowledge bots /RAG	L4 / A10 embeddings · L40S / A100 serving · H200 long context	**SPOT for ingestion & indexing · ON-DEMAND for serving One workload, two phases, two procurement models**	One workload, two phases, two procurement models
AI agents	L4 / L40S inference · A100 / H100 heavy reasoning pipelines	**ON-DEMAND production · SPOT async /background jobs**	GPU sits at the inference endpoint, not the wholeworkflow — external APIs may need none
AI coding assistants	L4 / A10 quantized · L40S / A100 / H100 large context, many devs	**ON-DEMAND — developers notice latency**	Context window x concurrent users drives VRAM, not model size alon
Document processing	T4 / L4 / A10 OCR & classification · L40S / A100 big batches, VLMs	**SPOT for batch ON-DEMAND only if near-real-time**	Bottleneck is often I/O and the task queue, not the GPU
Batch inference at scale	L4 / A10 cost-efficient · L40S heavier · A100 / H100 if time-boxed	**SPOT, almost always**	Optimize throughput per euro; queue, checkpoint, retry, shut dow
Computer vision	L4 / A10 inference · L40S / RTX 6000 graphics · A100 / H100 training	**ON-DEMAND / edge for live streams · SPOT, for archive & training**	Video streams: NVENC/NVDEC throughput beats raw compute
Speech AI / call analytics	L4 / A10 real-time · T4 batch transcription · A100 / H100 high volume	**ON-DEMAND live · SPOT recorded**	Live = latency budget; archives = cost per processed
Synthetic data generation	L40S / RTX 6000 media (48 GB) · A100 / H100 / H200 heavy generation	**SPOT, almost always Batch**	Batch by nature; VRAM caps batch size for diffusion / video
Recommendation / fraud /risk	L4 / A10 scoring · A100 / H100 training deep recommenders	**ON-DEMAND serving · SPOT training & batch scoring**	In-flow checks can't wait; classic ML may not need a GPU at all
Gaming / media	AI L40S / RTX 6000 / A10 assets & rendering · L4 transcode A100/ H100 training	**ON-DEMAND interactive · SPOT batch generation, render-like jobs**	Graphics / media engines matter more than an H100's compute
HPC / scientific simulation	A100 / H100 / H200 heavy compute · L40S mixed simulation + viz	**SPOT if checkpointed · RESERVED / ON-DEMAND otherwise**	Restart cost decides the market, not the hourly rate

The card ladder

VRAM · memory bandwidth · role

T4	**16 GB · ~320 GB/s** dev / test · cheap validation
L4	**24 GB · ~300 GB/s** prod default · NVENC/NVDEC video
A10	**24 GB · ~600 GB/s** same VRAM as L4, ~2x decode speed
L40S	**48 GB · ~864 GB/s** large-model serving · training workhorse
A100	**40/80 GB · 1.6-2 TB/s** large-model serving · training workhorse
H100	**80 GB · ~3.3 TB/s** throughput ceiling
H200	**141 GB · ~4.8 TB/s** VRAM ceiling: long context, big batches

Does the workload fit?

Weights: params x 2 B (FP16) or ~0.5 B (INT4).
8B = ~16 GB · 70B = ~140 GB unquantized -
a 70B does NOT fit one A100 80GB.
KV cache: ~0.5 MB / token (7-8B class).
10 chats x 4k tokens = ~20 GB on top of weights.
Capacity per card = (VRAM - weights) / KV per request.

Memory picks the class

Weights + KV cache, not TFLOPs. “The model fits” and “the workload fits” are different claims.

Bandwidth picks tokens/sec

Decode re-reads the weights every token. Benchmark first-token and inter-token latency separately, under load.

Concurrency picks scale-up vs out

Continuous batching (vLLM/TGI) multiplies throughput - and KV cache. Size from concurrent requests.

Interruption tolerance picks the market

GPU spot: typically 30-50% below on-demand, reclaimed on 30 s - 2 min notice. High-demand cards (H100) often have little or no spot at all.