1. GPU Runtime State (Critical)
The tool needs actual GPU metrics from running pods, not just K8s resource requests. This requires:

Execute nvidia-smi via kubectl exec inside pods
Extract: GPU type, memory used/total, utilization %, temperature
Handle cases where nvidia-smi isn't available gracefully

Can you add this?

2. Inference Runtime Configuration (Critical)
We need ML-specific config details:

Runtime engine (vLLM, Triton, TGI) and version
Precision (fp16, int8, etc.)
Batch size, max context length
Tensor parallelism settings

These come from container args, env vars, and potentially runtime APIs. Can you parse these for common frameworks?

3. Model Identification (Important)
Beyond deployment name, we need:

Actual model name (e.g., from MODEL_NAME env var or --model arg)
Model source/registry path
Architecture and parameter count if detectable

Can you add model metadata extraction?

4. Multi-Pod Aggregation (Important)
For deployments with multiple replicas, should output be:

One aggregated ModelSpec with replica count and avg metrics? (preferred)
Individual reports per pod?

Please clarify your approach.

5. Output Formats (Important)
Can you support:

YAML (default)
JSON (--format json)
One-line summary (--format oneline)

6. Error Handling (Important)
If nvidia-smi fails or a pod is unreachable, should the tool:

Mark fields as "unknown" and continue? (preferred)
Fail the entire report?

7. Issues Detection (Nice to have)
Can you add an "issues" section that flags:

GPU memory under-utilized
Missing resource requests
Unhealthy replicas
Common misconfigurations

8. Testing (Important)
Can you include:

Unit tests with mocked K8s responses
Integration tests against kind/minikube with sample ML workloads