SPACSEC 2026 · SAINT PETERSBURG · 26 OCTOBER 2026 · SESSION 3 · PAPER #22

Context-Instrumental Data Distillation for Kubernetes Manifest Generation Method and Experimental Evaluation

Can a 1.5B model fine-tuned on a laptop CPU write Kubernetes manifests that pass strict validators?

Andrey V. Kozachok · Anatoliy A. Bakaev · Aleksandr V. Kozachok · Shamil Magomedov · Artem N. Noev

Institute of Cybersecurity and Digital Technologies, MIREA – Russian Technological University, Moscow

MOTIVATION

Local generation for protected infrastructure

Operational risk

- Linked Deployment, Service and RBAC objects
- Wrong fields can break deployment
- Excess privileges create security risks

Critical infrastructure

- Protected configurations must stay inside
- External APIs may be prohibited by security policy
- Synthetic tasks can go to the teacher

Small models (1–4B)

- Local inference on a laptop CPU
- No GPU required in this pilot
- Training pairs pass automated domain checks

A local assistant can use synthetic training data while real configurations stay inside the protected perimeter.

METHOD

Training pairs pass domain checks before fine-tuning

Teacher LLM

generates task–manifest pairs

Validators L1–L4

keep only pairs passing configured checks

Student SLM

fine-tuning on pairs that pass the checks

Ordinary supervised fine-tuning

The teacher provides examples.
Teacher logits are not used.
The filter changes which examples the student sees.

Checks cover selected properties.
Passing them does not guarantee full correctness.

An incomplete Role rule

```
rules:
- apiGroups: [""]
  resources: ["pods"]
  # verbs is missing
```

Valid YAML, rejected at L2
Illustrative fragment

CONTEXT MODEL

Shared domain context for generation and validation

$CM = \langle M, G, S, P, V \rangle$

M — Meta-spec

allowed apiVersion + kind

G — Grammar

YAML syntax

S — Schema

OpenAPI types, required fields

P — Compositions

Deployment → Service → Ingress

V — Validators

$v(a) \in \{0, 1\}$

↓ *checked as four successive levels*

L1
YAML syntax
from G

L2
Strict schema
from M, S

L3-lite
Cross-resource links
from P

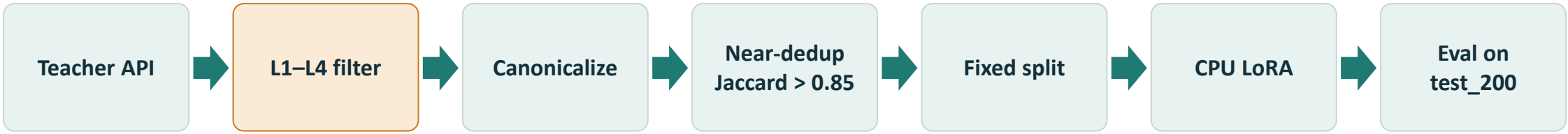
L4
Security policies
critical rules

Used twice. Fragments of CM go into the teacher prompt, and CM defines what the filter checks.

PIPELINE

Four validation levels, then deduplication and a fixed split

Level	Tool	What it checks
L1	YAML parser	YAML syntax and output format
L2	kubeconform --strict Kubernetes 1.30.0	Required fields, types and API versions
L3-lite	2 cross-resource checks	Service selectors and HPA targets
L4	7 critical Checkov rules	Resource requests / limits and container privileges



Other Checkov / Trivy findings are warnings. No server-side kubectl dry-run.

SETUP

A cloud teacher, a 1.5B student and a laptop CPU

First teacher batch

182 / 1,613 rejected

1,431 accepted
About 1 in 9 responses rejected

Final corpus of 1,710 unique records

Train 1,200 Validation 100 Test 200
210 records outside this final split
18 resource types in test_200

Teacher

DeepSeek-V4 Flash via API
Synthetic tasks only

Student

Qwen2.5-Coder-1.5B-Instruct
LoRA r = 4, α = 8 · 1 epoch · fp32

Hardware

Core Ultra 7 255H · 32 GB RAM
No GPU · one run per setting · seed 42

Generation settings raised full-pass@1 from 82% to 91%



+9.0 pp

Same 1K adapter, no retraining. Prompt tightened and token limit raised from 512 to 768 together.

183 / 200

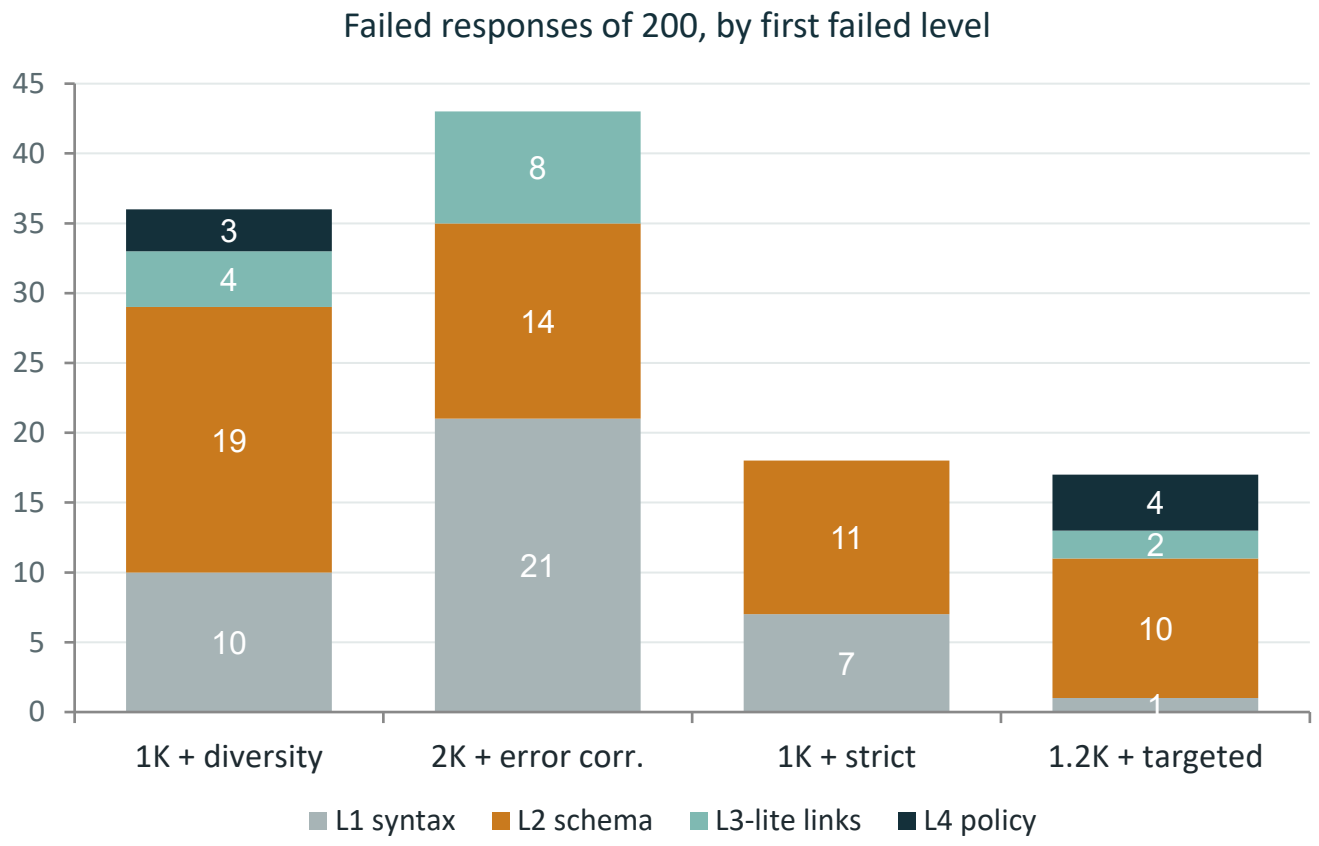
Best observed score on the synthetic development set. Wilson 95% CI 86.8–94.6%.

The 2K run used a different corpus build. This is not a controlled data-size comparison.

1 run · seed 42

RESULT 2 · WHERE OUTPUTS FAIL

No confirmed gain from 91.0% to 91.5%



91.0% vs 91.5%

200 paired tasks
171 both pass · 6 both fail
11 first only · 12 second only
Exact McNemar p = 1.0

No evidence of improvement

Paired McNemar test gives p = 1.0.

Wilson CIs 86.2–94.2% and 86.8–94.6%.

CIs cover task sampling, not seeds

RESULT 3 · BEST RUN IN DETAIL

StatefulSet and RBAC rules remain the weak spots

Resource type	Pass / N	full-pass@1
StatefulSet	8 / 15	**53.3%**
ClusterRole	11 / 14	78.6%
Job	9 / 11	81.8%
ClusterRoleBinding	8 / 9	88.9%
Ingress	10 / 11	90.9%
Role	11 / 12	91.7%
Deployment	33 / 35	94.3%
Other 11 types	93 / 93	100%

Typical failures

- **StatefulSet** volumeMounts misplaced (5); selector ≠ Pod labels (2)
- **Role / ClusterRole** required rules[].verbs missing
- **Job** 2 outputs run as root → critical L4 policy

Cost on CPU

≈ 5 h

training (302 min)
peak RAM 14.6 GB

≈ 3 h

200 tasks (181 min)
~54 s per task on average

100% in small subgroups is not a stable quality estimate.

LIMITATIONS

Limits of the pilot evaluation

Development benchmark

Test IDs excluded and near-duplicates removed. Aggregate test errors guided later batches.

No matched controls

No matched unfiltered SFT, base-model or teacher evaluation. Filtering’s separate effect needs a matched comparison.

One run per setting, seed 42

Intervals describe task outcomes, not variation across seeds.

Partial checks

2 links and 7 security rules. No live API check or full verification of user intent.

Next steps — matched controls, several seeds and an independent benchmark

TAKEAWAYS

Practical benefits of the pilot

1. Real configurations stay inside the protected perimeter
2. Student training and inference run on a laptop CPU
3. Automated selection replaces manual checks of covered properties
4. +9.0 pp on the same adapter, without retraining
5. 183 / 200 pass configured checks on a synthetic development set

Generated manifests remain drafts for verification

Corpus, predictions, adapters, scripts · doi.org/10.5281/zenodo.23000787

Thank you! Questions?

Selected approaches and the role of validators

Prior work	Role of checks
CloudEval-YAML, IaC-Eval	Tests evaluate generated output
Minna et al. (Helm)	Analyzers guide repair
Deé-Lukács et al.	Template filling with schemas + ASP
This work	Validators filter the training corpus

The implementation uses ordinary supervised fine-tuning on filtered teacher examples. No teacher logits.

Validation yield of teacher-generation batches

Batch	Input	Accepted	No YAML	L1	L2	L3	L4
Balanced 1K	1,613	1,431	68	28	38	8	40
Diverse top-up	791	733	33	0	25	0	0
2K error-aware	1,183	1,131	30	3	19	0	0
Targeted corr.	293	264	11	1	4	1	12

Rejections use the first rejection reason. These are generation batches, not a teacher score on test_200.

Final split uses 1,200 + 100 + 200 = 1,500 of 1,710 records. The other 210 were not included.