Multi-Model Synthetic Training for Mission-Critical Small Language Models

Nolan W. Platt[1] Pragyansmita Nayak[2]

[1]Department of Computer Science
Virginia Tech
Blacksburg, VA

[2]Chief Data Scientist
Hitachi Vantara Federal
Reston, VA

3rd International Conference on Foundation and Large Language Models
November 2025
Vienna, Austria

Table of Contents

Table of Contents

Defn: Mission-Critical

Broadly speaking, 'mission-critical' refers to any system that **cannot fail**. In the United States, this typically refers to systems used in (1) defense, (2) intelligence, and (3) national security.

LLMs in Mission-Critical

With the rise of LLMs, there are several factors that have made mission-critical LLMs particularly challenging:

- Air-gapped (classified) environments.

LLMs in Mission-Critical

With the rise of LLMs, there are several factors that have made mission-critical LLMs particularly challenging:

- Air-gapped (classified) environments.
- Lack of abundance of training data (domain-specific)

LLMs in Mission-Critical

With the rise of LLMs, there are several factors that have made mission-critical LLMs particularly challenging:

- Air-gapped (classified) environments.
- Lack of abundance of training data (domain-specific)
- Cost of both inference and training such models.

AIS Data

Our paper focuses on developing a model for maritime intelligence – one that can answer natural language queries about tens of thousands of vessels throughout U.S. waters.

We chose this domain as we have an abundance of Automatic Identification System (AIS)[1] data – broadcasted every few minutes from transceivers on every vessel in our world.

Yet, we lack any labeled or otherwise pre-processed training data.

[1] AIS data contains useful information like speed over ground (SOG), course over ground (COG), lat/lon, vessel type, cargo, and souls onboard

Research Questions

Our paper seeks to answer the following research questions:

1. How do we turn billions of raw AIS transceiver data, formatted in CSV files, into useful Q&A training data for a language model?

Research Questions

Our paper seeks to answer the following research questions:

1. How do we turn billions of raw AIS transceiver data, formatted in CSV files, into useful Q&A training data for a language model?
2. How do we prevent model collapse and overfitting from our synthetic data?

Research Questions

Our paper seeks to answer the following research questions:

1. How do we turn billions of raw AIS transceiver data, formatted in CSV files, into useful Q&A training data for a language model?
2. How do we prevent model collapse and overfitting from our synthetic data?
3. Can specialized small language models (SLMs) be used in mission-critical environments, and if so, are they both trustworthy and affordable?

Table of Contents

AIS Data Sampling and Processing

3.2 billion AIS records of raw AIS data were provided by the U.S. Coast Guard, covering all tranceiver data from FY2024.

AIS data was then split into contexts, with each containing $[200, 500]$ vessels with complete positional data.

Sampling

Geographic regions (East Coast, West Coast, Gulf of Mexico, Great Lakes), ports, open water, diverse time periods, vessel types, and traffic densities.

Synthetic Q&A Generation Pipeline

We use DataDreamer (Penn) to generate 21,543 Q&A pairs via a multi-model approach. 90% of the pairs were used for training, while 10% were reserved for validation. Each context generated 12 questions across six diverse categories:

1. Trajectory predictions
2. Movement analysis
3. Vessel counting
4. Data analysis
5. Pattern detection
6. Anomaly detection

Preventing Model Collapse

Models trained on synthetic data from a *single* LLM can inherit that model's biases and limitations.

Different models exhibit different generation patterns and problem-solving approaches. To prevent overfitting, we alternated between two models every seven contexts:

GPT-4o (85.7%)

Strong at probabilistic trajectory predictions

o3-mini (14.3%)

Focused on rule-based violations and thresholds

Result: Model generalizes across reasoning styles (75.9% vs 71.4% accuracy on respective test sets).

Model Selection

Magistral Small (7B)

Overfit; memorized trivial patterns without comprehension.

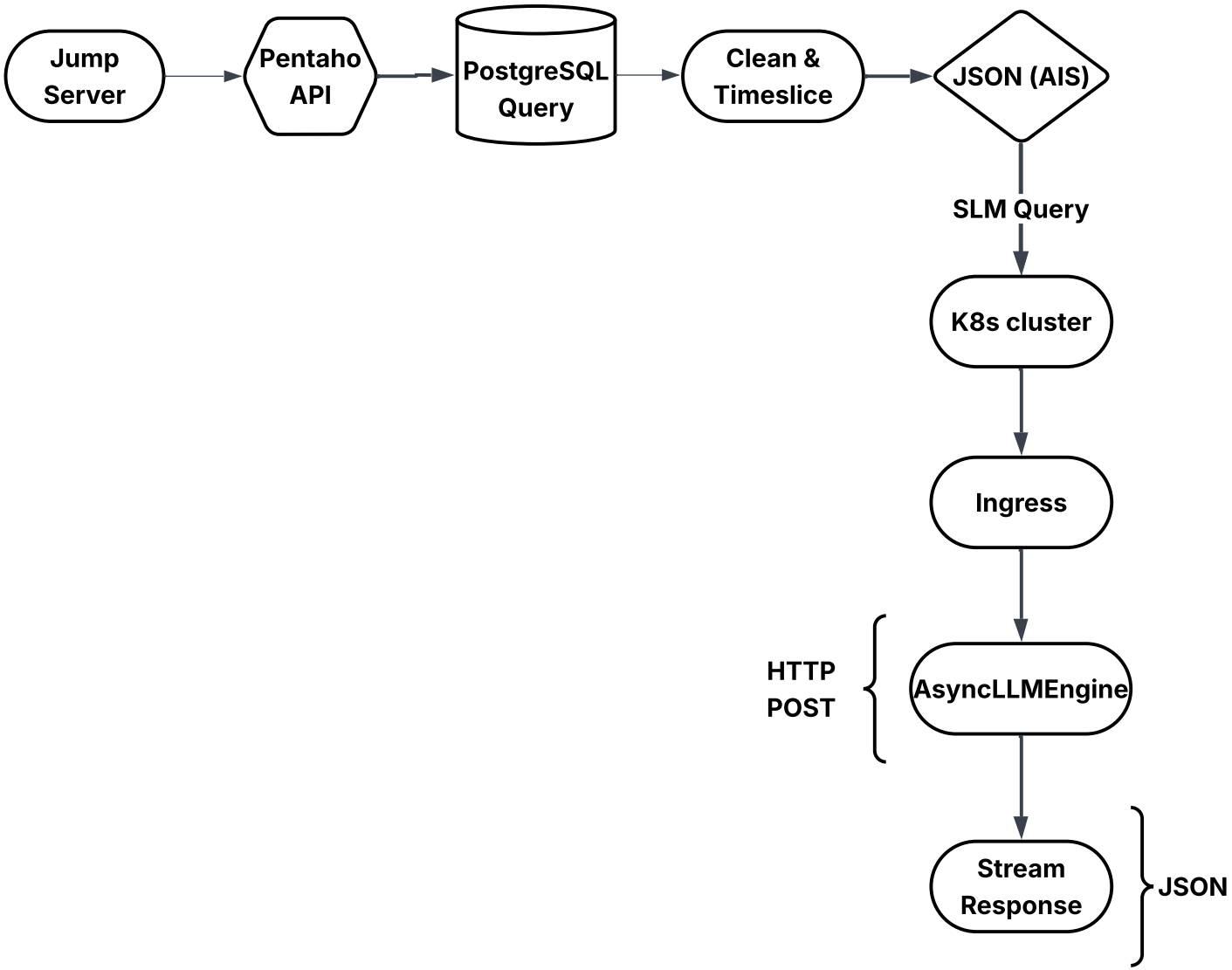
Llama 3.1 (8B)

Constant hallucination of vessel positions.

Qwen2.5 (7B)

Final model used. Selected for JSON pre-training and native long-context support via YaRN scaling.

System Architecture



Context Extension via YaRN Scaling

$$h(\theta_d) = (1 - \gamma(r(d))) \frac{\theta_d}{s} + \gamma(r(d))\theta_d$$

- Let scale factor $s = 4$, which extends our context window from 32k to 131k tokens (a 4x increase)
- $\gamma(r(d))$ smoothly transitions from 0 to 1 based on the frequency's wavelength
- $r(d) = L/\lambda_d$, where L is the original context length and $\lambda_d = 2\pi b^{2d/|D|}$ is the wavelength at dimension d
- $\theta_d = b^{-2d/|D|}$ represents the original RoPE frequency at dimension d, with base $b = 10000$

Table of Contents

Evaluation Methods

We evaluate our model via both standard NLP metrics and domain-specific metrics.

NLP

BLEU, ROUGE-L, BERTScore F1

Domain

Manual accuracy, shows reasoning (Y/n), avg. response length

The Evaluation Paradox

NLP Metrics

BLEU	0.09%
ROUGE-L	10.9%
BERTScore F1	-0.18

Assessment: extremely poor.

Domain Metrics

Manual Accuracy	75.0%
Shows Reasoning	98.0%
Automated (n=500)	70.8%

Assessment: very strong performance.

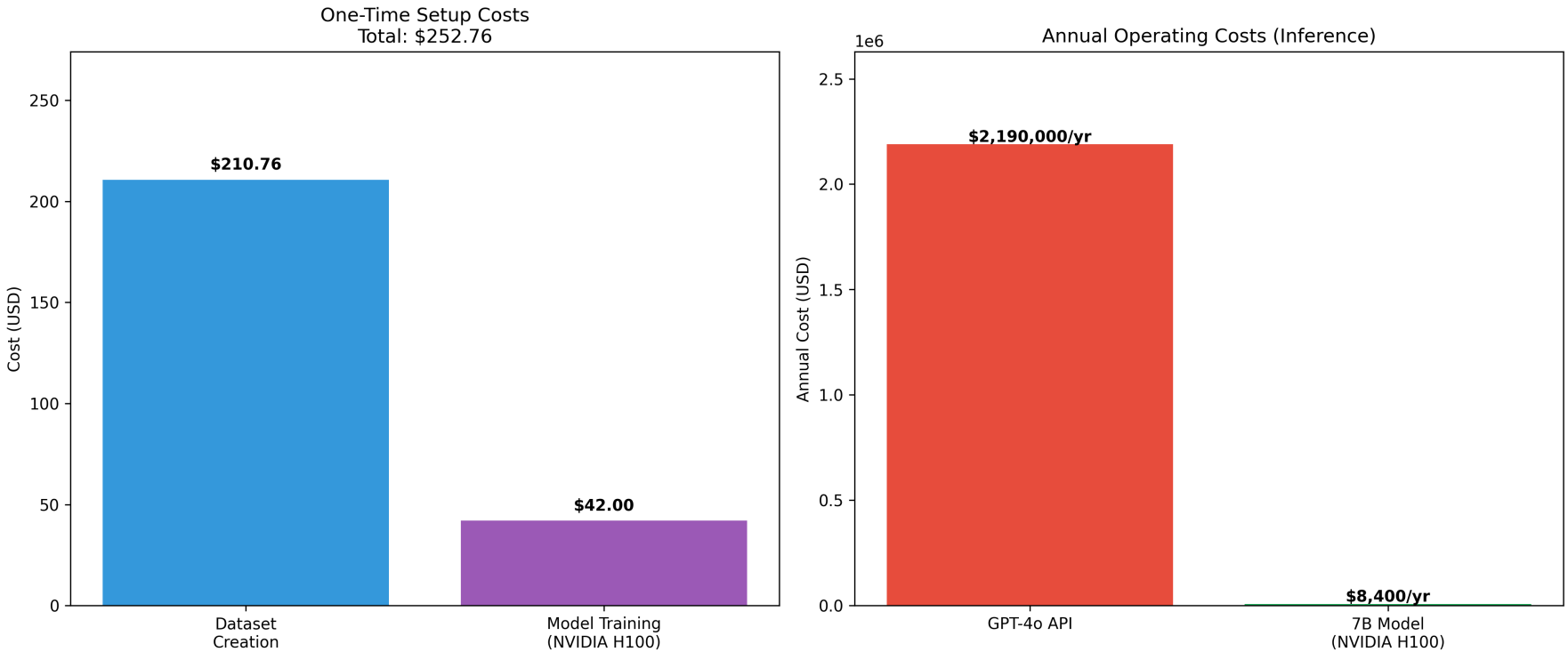
Why the gap? Our model produces verbose, educational responses (9.2x longer than reference answers). BLEU penalizes this; humans value it.

Performance by Question Type

Category	Accuracy	95% CI
Anomaly Detection	100%	[64.6%, 100%]
Trajectory Prediction	81.5%	[63.3%, 91.8%]
Pattern Detection	83.3%	[55.2%, 95.3%]
Vessel Counting	70.6%	[46.9%, 86.7%]
Data Analysis	65.2%	[44.9%, 81.2%]
Movement Analysis	61.5%	[35.5%, 82.3%]

$\therefore$ Model excels at clear threshold violations (anomalies), struggles more with nuanced heading/acceleration interpretation.

Cost Analysis: 261x Reduction



Calculations based on ∼10,000 queries/day.

GPT-4o: $2.19M/year.

Ours: $8,400/year (single H100).

Limitations

- **Temporal**: Model trained on 2024 data; maritime patterns evolve
- **Geographic**: U.S. waters only; international deployment needs region-specific tuning
- **Adversarial**: Cannot catch all AIS spoofing – high-stakes needs human oversight
- **Context**: Peak-hour major ports may exceed 131k token window

Conclusion

What we showed:

1. First maritime intelligence dataset for language models **in the world**
   - 3.2B AIS records $\rightarrow$ 21,543 Q&A pairs
2. Multi-model synthetic generation prevents overfitting
3. SLMs can handle mission-critical tasks: 75% accuracy, 261x cheaper
4. Traditional NLP metrics don't capture domain-specific performance

Future Directions

Neurosymbolic AI (Scallop), agentic systems, ecosystem of specialized SLMs.

Use expensive LLMs as teachers, not workers.

One-time synthetic generation $\gg$ continuous expensive inference

Dataset, code, model publicly available:
`huggingface.co/nolanplatt/hvf-slm-v3-qwen`

Questions?
nolanplatt@vt.edu