Agentic Knowledge Distillation

Autonomous Training of Small Language Models for SMS Threat Detection

Adel ElZemity · Joshua Sylvester · Budi Arief · Rogério de Lemos

School of Computing, University of Kent, Canterbury, UK

AI-SS 2026

Content

- Motivation
- Proposed Approach
- Experimental setup
- Experiments
- Implication and Limitations
- Conclusions
- Future Work

SMS Phishing is Surging

+307%

Surge in smishing attacks in recent years [1]

High Trust Channel

SMS bypasses email filters. Users inherently trust text messages more than email, making click rates far higher

Fast-Evolving Tactics

Attackers use URL shorteners, homoglyphs, and social engineering. Static rule-based filters fail quickly

Privacy Constraints

Cloud-based filtering exposes private messages to third parties making on-device detection essential

[1] https://www.robokiller.com/robokiller-2022-phone-scam-report

Why Not Just Use a Large Language Model Directly?

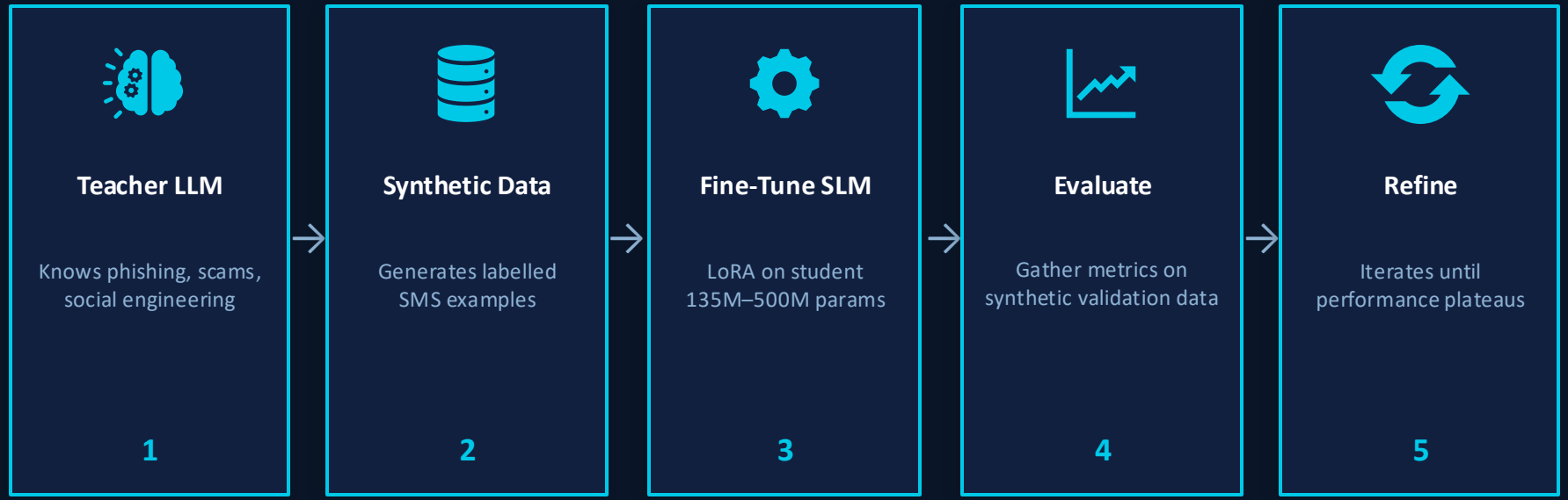
Large Language Models (LLMs)

- Billions of parameters meaning it cannot run on a phone
- High latency which is unacceptable for real-time SMS filtering
- Privacy risk since messages leave the device
- Expensive API calls per message

Small Language Models (SLMs)

- 135M-500M parameters therefore it runs on small devices
- Fast on-device inference leading to real-time filtering
- Private since messages never leave the device
- One-time fine-tuning cost, zero API calls

Proposed Approach - Agentic Knowledge Distillation



Agentic Knowledge Distillation Closed-Loop

1

Task Spec Given to Teacher LLM

Identical system prompt across all 4 teacher LLMs: mission, constraints, iteration structure.

2

Generate Synthetic Training + Validation Data

2,000+ synthetic SMS (50/50 spam/ham). Validation set is fixed at 500 samples.

3

Fine-Tune Student SLM with LoRA

LoRA rank 32, LR 5×10^{-5}, batch size 8. Only small matrices updated

4

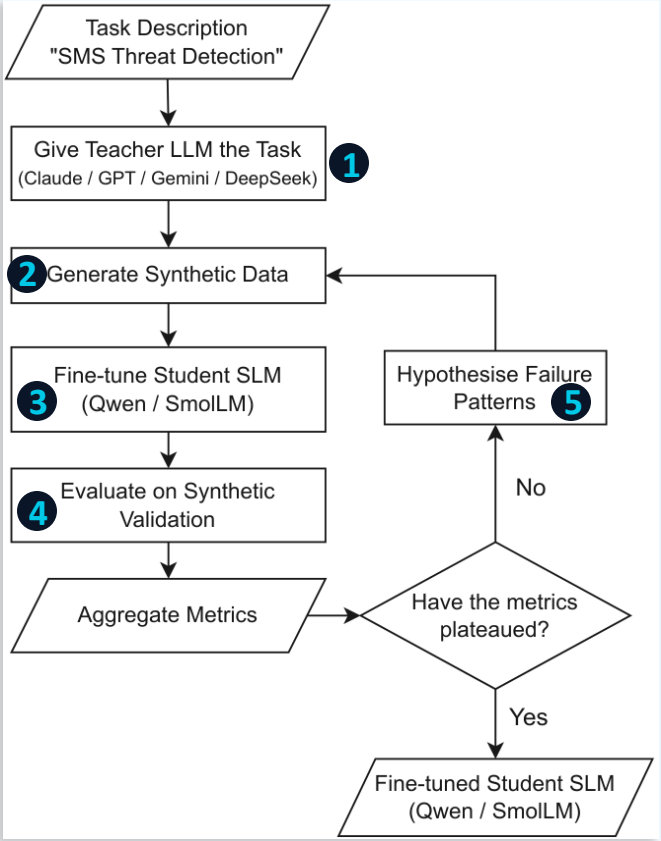
Evaluate on Synthetic Validation Set

Metrics: Accuracy, Precision, Recall, FP rate, FN rate fed back to teacher.

5

Hypothesise Failure Patterns & Refine

Teacher analyses aggregate metrics, generates hard negatives, loops until plateau.



Experimental Setup

Teacher LLMs

Claude Opus 4.5
Anthropic

GPT 5.2 Codex
OpenAI

Gemini 3 Pro
Google

DeepSeek V3.2
DeepSeek

Student SLMs (on-device, consumer hardware)

Qwen2.5-0.5B-Instruct
494M parameters

SmolLM2-135M-Instruct
135M parameters

Evaluation Dataset & Conditions

- **SMS Spam Collection:** 5,574 real messages [2]
- **Balanced test subset:** 1,494 messages — 747 spam, 747 ham

[2] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, “Contributions to the study of SMS spam filtering: new collection and results,” in Proceedings of the 11th ACM Symposium on Document Engineering (DocEng ’11), Mountain View, CA, USA, 2011, pp. 259–262.

Experimental Setup

Baselines

Zero-shot

Testing without any task-specific training, relying only on its pre-trained knowledge and the given prompt

Direct Preference Optimisation (DPO)

Testing a model after using a method that trains a model to prefer preferred outputs over rejected ones

Proposed Method

Agentic Knowledge Distillation (KD)

Testing a model after it has undergone the iterative process previously described

Baseline Results - Zero-shot & DPO Fine-Tuning

Zero-Shot (No Fine-Tuning)

Qwen2.5-0.5B
Acc: 49.80% | Recall: 0.27%

Almost all predictions: 'ham'

SmolLM2-135M
Acc: 46.85% | Recall: 45.11%

Near random chance

DPO with Teacher-Generated Synthetic Data (10,000 samples, single pass)

Teacher LLM	Student SLM	Accuracy	Precision	Recall
Claude Opus 4.5	Qwen2.5-0.5B	52.74%	72.52%	52.74%
	SmolLM2-135M	**80.25%**	**81.81%**	**80.25%**
GPT 5.2 Codex	Qwen2.5-0.5B	51.34%	57.30%	51.34%
	SmolLM2-135M	52.61%	70.57%	52.61%
Gemini 3 Pro	Qwen2.5-0.5B	52.54%	68.93%	52.54%
	SmolLM2-135M	68.27%	77.33%	68.27%
DeepSeek V3.2	Qwen2.5-0.5B	52.48%	68.09%	52.48%
	SmolLM2-135M	76.44%	78.44%	76.44%

Agentic Knowledge Distillation Results

Teacher LLM	Student SLM	Accuracy	Precision	Recall	F1	Time
Claude Opus 4.5	Qwen2.5-0.5B	**94.31%**	92.65%	**96.25%**	**94.42%**	~7 min
	SmolLM2-135M	**86.28%**	80.38%	95.98%	**87.00%**	~6 min
DeepSeek V3.2	Qwen2.5-0.5B	92.10%	91.77%	92.50%	92.13%	~8 min
	SmolLM2-135M	86.21%	**88.70%**	83.00%	85.75%	~7 min
Gemini 3 Pro	Qwen2.5-0.5B	85.21%	79.29%	95.31%	86.58%	~9 min
	SmolLM2-135M	80.46%	72.73%	97.46%	83.31%	~8 min
GPT 5.2 Codex	Qwen2.5-0.5B	71.08%	**98.76%**	42.70%	59.64%	~6 min
	SmolLM2-135M	59.71%	55.38%	**99.87%**	71.26%	~5 min

+44.51pp

Best Accuracy Gain

Claude + Qwen

99.87%

Best Recall

GPT + SmolLM2

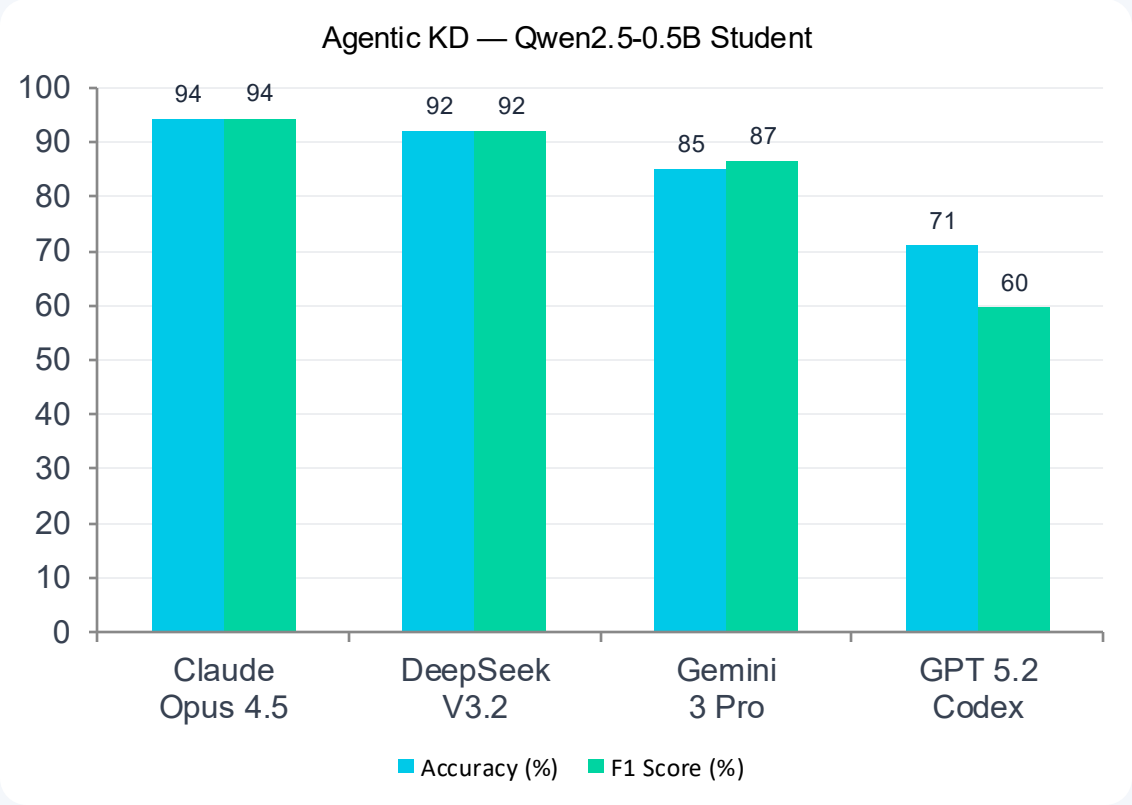
(but 55% precision)

94.42% F1

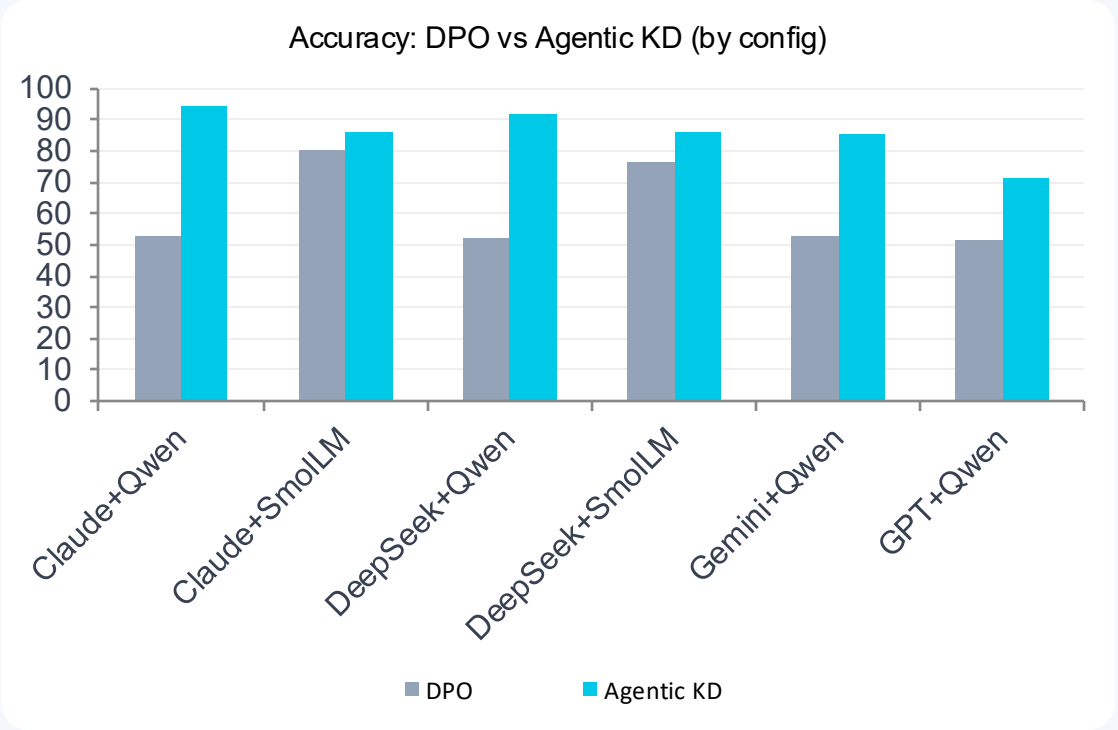
Most Balanced

Claude+ Qwen

Teacher LLM Performance Comparison



Agentic KD compared to DPO



Discussion: Implications & Limitations

Security Implications

Label Scarcity Paradox Solved

No labelled threat data needed. LLM's knowledge of phishing, scams, and social engineering bootstraps the detector

Privacy

On-device inference means personal messages never touch external APIs. Perfect for SMS which carries sensitive content

Adaptability

Re-run the pipeline in <10 minutes. No data collection, no manual annotation

Limitations

Bounded by teacher knowledge

Emerging attack patterns absent from LLM training data may be missed entirely

Synthetic validation bias

Internal loop is guided by teacher-generated 'ground truth' meaning there could be a possible distributional mismatch to real-world SMS phishing

Conclusions and Future Work

1

Agentic KD works

LLMs can autonomously fine-tune SLMs for security tasks with no labelled data, no human engineers, <10 minutes

2

Teacher LLM choice is critical

25+ pp performance gap between best (Claude: 94%) and worst (GPT: 71%) teacher. Not all LLMs are equal ML engineers

3

Closed-loop feedback is key

DPO achieves 50–80% accuracy. Agentic KD achieves 60–94%. Iterative error-driven refinement makes the difference

4

On-device security is viable

96.25% recall on real SMS spam. Privacy-preserving, consumer-hardware deployable

5

Single teacher per run

Ensemble of teacher could combine strengths of different models

Thank You!

Questions

Joshua Sylvester

jrs71@kent.ac.uk

School of Computing, University of Kent, Canterbury, UK