DeEscalWild

A Real-World Benchmark for Automated De-Escalation Training with SLMs

Md Hasebul Hasan, Krity Haque Charu, Eshwara Prasad Sridhar,
Shuchisnigdha Deb, Mohammad A. Islam
University of Texas at Arlington

NeurIPS 2026, Evaluations & Datasets Track

Code

Data

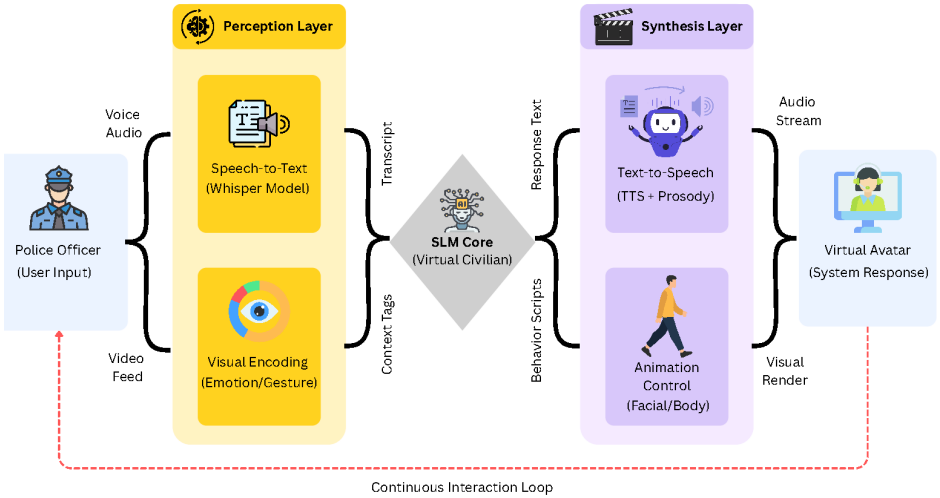
What Is De-Escalation and Why It Matters

- De-escalation means calming a tense situation by talking, not by using force
  - stay calm, listen, show you understand, give the person time and space
- Police use it in many situations: traffic stops, family fights, mental health crises
- Police meet people very often
  - about 54 million people in the US had contact with police in 2020
- Training helps. After Louisville police took de-escalation training:
  - 28% less use of force
  - 26% fewer people hurt, and 36% fewer officers hurt
- But it is a skill, so officers need a lot of practice

BJS, Contacts Between Police and the Public, 2020. Engel et al., Criminology & Public Policy, 2022.

Motivation

- How officers practice today
  - live role-play: expensive, hard to repeat consistently
  - branching videos: scripted, only a few fixed paths
- Idea: the officer talks, an AI civilian responds in VR
- Problems with large LLMs
  - too big to run on a VR headset or edge device
  - prior prototype took >4 s per reply
  - closed models can't be fine-tuned for this domain



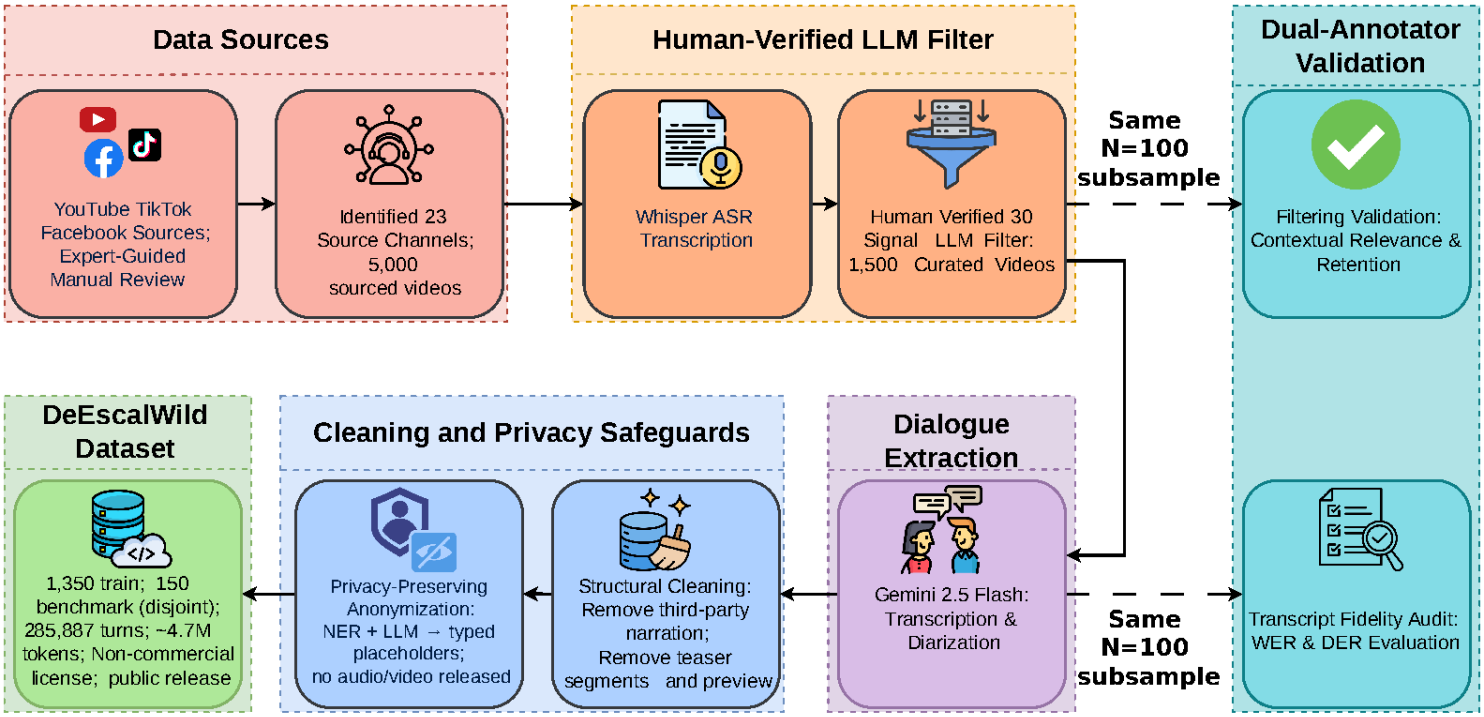
Our envisioned training system (Fig. 6). This paper builds the SLM civilian.

Sridhar et al., 2025. Anand & Polyak, 2024.

Why Small Language Models?

- SLMs (2-3B parameters) run fast and on-device
  - sub-second replies, no internet connection needed
  - can be fine-tuned on one consumer GPU
- But out of the box they don't behave like real civilians
  - they sound like polite assistants
- They need domain-specific training data
  - and there is no public dataset of real de-escalation dialogue

Data Curation Pipeline



5,000 public videos → 1,500 clean, anonymized, speaker-labeled transcripts

Privacy and Release

- We share only the text of the conversations
  - no videos, no audio, no links to the original videos
- We hide personal details
  - names become [CIVILIAN_NAME], places become [LOCATION]
- We checked 150 transcripts by hand
  - after our final fix, only about 1% of personal details were missed
- Access is for research only, not for business use

Item	Released?
Text transcripts	Yes (gated)
Video, audio, thumbnails	No
Names, locations	No, replaced with labels
Source URLs, usernames	No
Code	Yes

Dataset Statistics

Statistic	Value
Scenarios	1,500
Dialogue turns	285,887
Words	3.63M
Tokens (est.)	4.71M
Avg. interaction	~18 min, ~190 turns
Split	1,350 train / 150 test

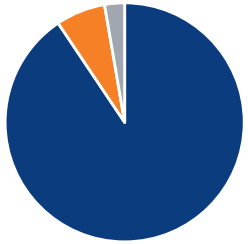
- Long conversations: models must track context over ~190 turns
- 90.6% of interactions are high-tension

Incident type



Public disturbance	29.4%
Suspicious activity	26.7%
Traffic stop	22.8%
Domestic disturbance	13.9%
Mental health crisis	6.3%
Other / mixed	0.9%

Tension level

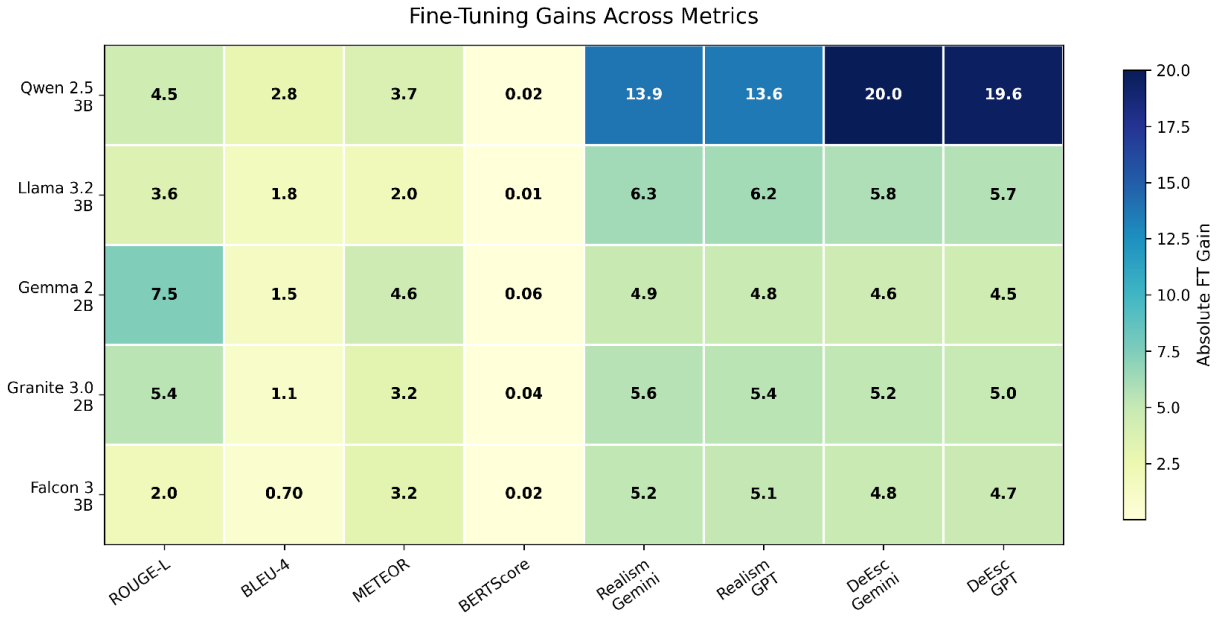


High	90.6%
Medium	6.7%
Other / mixed	2.7%

Experimental Set UP

- Model: Qwen 2.5 3B, Llama 3.2 3B, Gemma 2 2B, Granite 3.0 2B, Falcon 3 3B
- Baseline: Gemini 2.5 Flash with few-shot prompting (and with RAG)
- Evaluation Strategy: Pre-post fine-tuning evaluation protocol and compare with Gemini 2.5 Flash
- Metrics
  - lexical / semantic: ROUGE-L, BLEU-4, METEOR, BERTScore
  - realism score (0-100) from 4 LLM judges: Gemini 3.1 Pro, GPT-5.4, Llama 3.1 70B, Qwen 2.5 72B
  - blind human evaluation by 2 domain experts

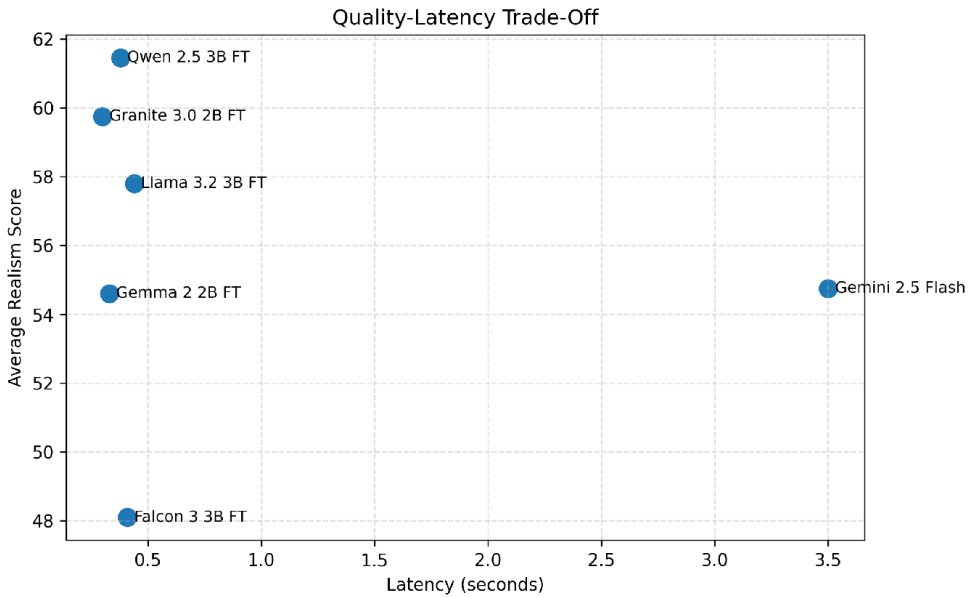
Results: Effect of Fine-Tuning



- Every model improves on every metric
- Largest gains on realism and de-escalation scores
- Realism (turn-level): +18 to +28 points for all 5 models
- Qwen 2.5 3B benefits the most
- Same trend under all 4 LLM judges

Absolute gain, base → fine-tuned (Fig. 3). Realism / De-esc columns from closed-loop simulation.

Results: Fine-Tuned SLMs vs. Gemini 2.5 Flash



Realism vs. latency (Fig. 4)

- Qwen 2.5 3B FT beats Gemini on all quality metrics
  - realism 62.1 vs. 55.4
  - 0.38 s vs. 3.50 s per reply
  - still better than Gemini + RAG (62.1 vs. 59.7)
- Human evaluation (2 experts, blind, out of 5)
  - Qwen 2.5 FT 4.25, Gemini 3.87, Llama 3.2 FT 3.63
  - base models at or below 2.30
  - mostly agrees with automatic metrics; Llama FT is the exception

Conclusion

- DeEscalWild: first large-scale benchmark from in-the-wild police-civilian interactions
  - 1,500 scenarios, 285K turns, anonymized text only
- Small fine-tuned models can beat a large general model on this task, at sub-second latency
- The curation pipeline plug and play in nature, can be reused for other domains
- Limitations and future work
  - we measure realism, not whether it improves officer performance
  - next: integrate into a VR trainer and run a user study

Code

Data

Thank you!

Questions?

mdhasebul.hasan@uta.edu