ConflLlama

Domain-specific Adaptation of Large Language Models for Conflict Event Classification

Shreyas Meher* Patrick T. Brandt†

* Erasmus School of Social and Behavioural Sciences, Erasmus University Rotterdam

† School of Economic, Political and Policy Sciences, The University of Texas at Dallas

September 12, 2025

The AI Accessibility Problem

- **Manual coding** is inconsistent and time-consuming
- **200,000+** events in Global Terrorism Database
- Events are becoming increasingly **complex**
- Need for **real-time**, **scalable** analysis

The Cost Barrier

State-of-the-art AI models are **expensive** and **inaccessible** to most researchers

The Cost of Cutting-Edge AI

Proprietary Models:

- **GPT-5:** \$1.25 / 1M tokens (8x for the output)
- **Claude 4.1 Opus:** \$15 / 1M tokens (5x for the output)
- **Both are reasoning models! They think too much and costs increase.**

Research Reality

Most political scientists have:

- **Limited budgets**
- **Basic hardware**

Vendor Lock-in

- **No control** over model updates
- **Rate limits** and availability issues
- **Data privacy** concerns

The Result

Cutting-edge AI remains locked away from the researchers who need it most

The Open Source Revolution

Open Source Models:

- **Llama 3.1:** **Free** to use
- **200K events:** **$0** in API costs
- **One-time training:** **$10-20**
- **Unlimited inference** after training

Democratized Access:

- **Hardware:** Consumer-grade GPUs
- **Training:** 1.5 hours on cloud
- **Deployment:** Laptop-friendly

Huggingface - ConflLlama

Training Dynamics

- **Training time:** 1.5 hours on H100 (also works on a 16 GB card)
- **Speed:** 3.49 seconds/iteration
- **Gradient norms:** Stable at 0.53

Accessibility

Consumer-grade hardware can fine-tune 8B parameter models

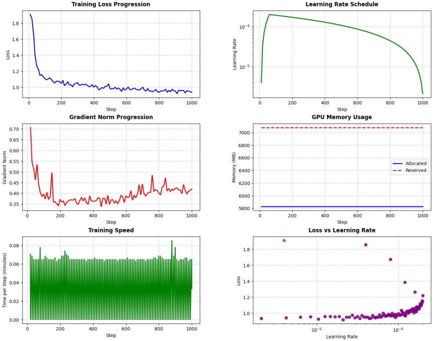
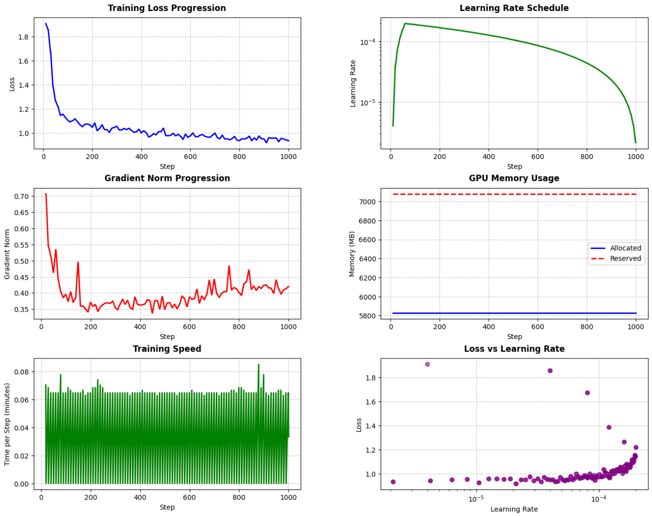


Figure: Training loss, learning rate, and resource utilization

Training Dynamics — Full View



Training loss, learning rate, and resource utilization

The Global Terrorism Database (GTD)

Dataset Characteristics:

- **Coverage:** 200,000+ events (1970-2020)
- **Scope:** Domestic & international terrorism
- **Attributes:** 120+ variables per event
- **Challenge:** Severe class imbalance

Our Focus:

- **Task:** Multi-label attack type classification
- **Labels:** 9 attack categories
- **Complexity:** 5.3% multi-label events

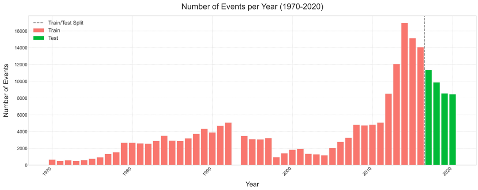


Figure: GTD events over time showing Train & Test split

Experimental Setup

- **Temporal Split:** Natural evaluation design
  - Training: 171,514 events (pre-2017)
  - Test: 38,192 events (2017 onwards)
- **Class Distribution Challenge:**
  - Bombing/Explosion: 48.7% of training data
  - Armed Assault: 23.6%
  - **Rare events:** Hijacking (0.4%), Unarmed Assault (0.5%)
- **Evaluation Focus:**
  - Multi-label classification metrics
  - Performance on rare event types
  - Temporal generalization capability

Why This Approach?

Why Llama 3.1?

- **Performance:** Matches GPT-4 on many tasks
- **Efficiency:** 8B parameters vs 175B+
- **Versatility:** Strong domain adaptation (Lu et al., 2024)
- **Context:** Extended context processing (128k with Unsloth edits)
- **Accessibility:** Open source & permissive license

Why QLoRA?

- **Memory Efficiency:** 16GB $\rightarrow$ 6GB (Dettmers et al., 2024)
- **Performance:** Minimal degradation from quantization
- **Speed:** Low-rank adaptation trains faster
- **Accessibility:** Consumer hardware deployment

Why Not BERT-based Models?

Previous Approaches:

- **ConfliBERT** (Hu et al., 2022)
- Strong on common event types
- Struggles with rare events
- Limited context understanding

LLM Advantages:

- **Context:** Better narrative understanding
- **Generalization:** Few-shot capabilities
- **Flexibility:** Multi-task learning

Evidence from Literature

- Ornstein et al. (2023): LLMs "significantly outperform existing automated approaches"
- Heseltine & Clemm von Hohenberg (2024): LLMs as substitute for human experts
- Our results: +1463% on rare events vs BERT

The Gap

No prior work on **efficient LLM adaptation** for conflict classification

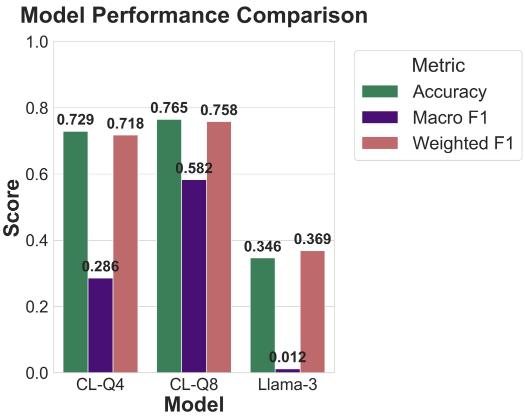
Dramatic Performance Improvements

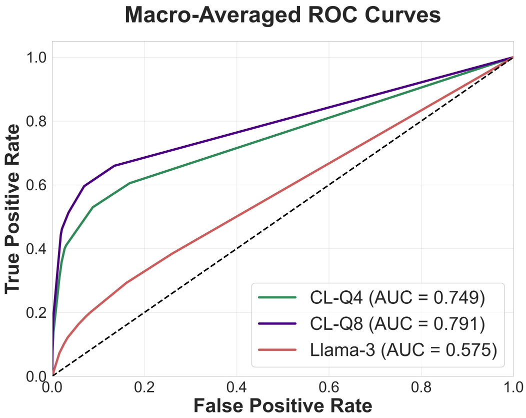
Attack Type	**Llama 3.1 F1**	**ConflLlama F1**	**Improvement**
Unarmed Assault	0.035	0.553	**+1463%**
Hostage Taking (Barricade)	0.045	0.353	**+692%**
Hijacking	0.100	0.629	**+527%**
Facility/Infrastructure	0.167	0.733	**+339%**
Assassination	0.201	0.655	**+226%**
Bombing/Explosion	0.549	0.908	**+65%**

Finding

Rare events show the most dramatic improvements - crucial for security applications

Model Performance Comparison





- CL-Q8 achieves highest overall performance across all metrics
- ROC curves show superior classification ability (AUC = 0.791)
- Llama-3 baseline comparison shows large improvement, even for Q4

Comparison with modernBERT

Attack Type	mBERT	ConflLlama
Overall Accuracy	**79.66%**	76.50%
Barricade	0.15	**0.35**
Unarmed Assault	0.31	**0.55**
Hijacking	0.37	**0.63**
Bombing	**0.94**	0.91
Kidnapping	**0.91**	0.84

Trade-off

ConflLlama: **Superior** on rare events

Security Focus

Rare events often have **disproportionate impact**

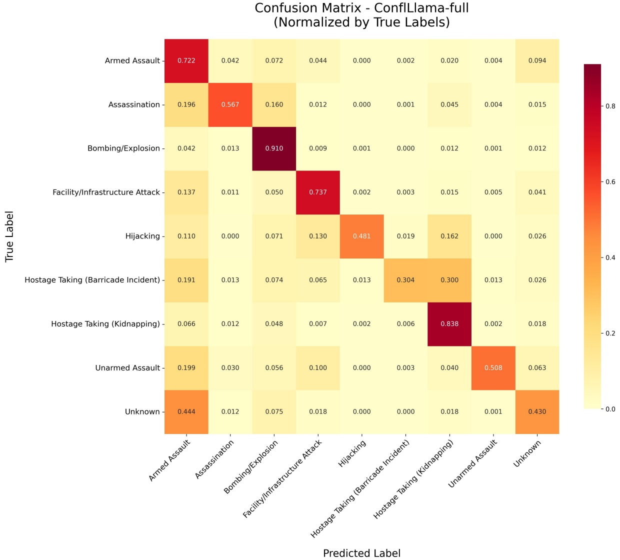
Both **modernBERT** and **ConflLlama** were finetuned on the *same training set and procedure*, ensuring fair comparison.

Multi-label Classification Excellence

- **Hamming Loss: 0.052**
- **Subset Accuracy: 72.4%** (exact match)
- **Partial Match: 73.8%**
- **Label Density:** 0.975 (true: 0.963)

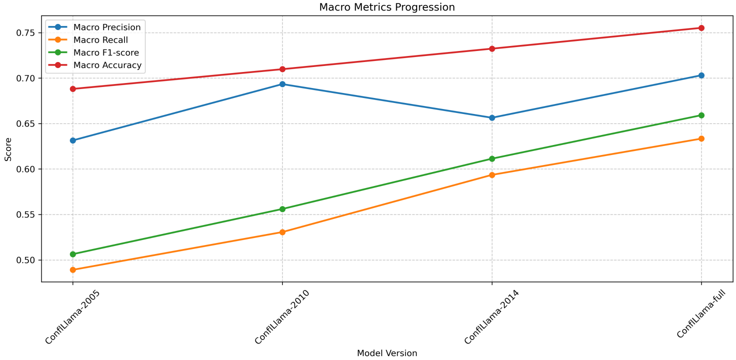
Real-world Impact

Mumbai, 2008: Armed Assault + Hostage Taking + Bombing



Temporal Coverage Impact

- **Model Accuracy:** 69% → 76% with expanded temporal data
- **Recall:** 49% → 63%
- **F1-score:** 51% → 66%



Democratizing Advanced NLP

- **Hardware Requirements: 16GB RAM** (consumer-grade)
- **Processing Speed: 44,280 events/hour**
- **Training Time: 1.5 hours** on cloud GPU
- **Memory Footprint:** $<$ **6GB** during fine-tuning

Research Impact

Makes cutting-edge conflict analysis accessible to researchers with **limited computational budgets**

Available: Huggingface - ready for immediate use

Key Takeaways

1. **Technical Innovation:** QLoRA enables **efficient adaptation** of large models
2. **Empirical Success:** Up to **1463% improvement** on challenging classifications
3. **Accessibility:** **Consumer-grade deployment** with professional results
4. **Robustness:** Consistent performance across different prompting strategies
5. **Multi-label Excellence:** Handles complex, overlapping conflict events

Bottom Line

ConflLlama **democratizes** advanced AI for political science while achieving **state-of-the-art performance**

Thank You

Questions & Discussion

Shreyas Meher
Erasmus University Rotterdam
meher@essb.eur.nl

Patrick T. Brandt
UT Dallas
pbrandt@utdallas.edu

Citation:

Meher, S., & Brandt, P. T. (2025). *ConflLlama: Domain-specific adaptation of large language models for conflict event classification*. Research & Politics, 12(3), 20531680251356282. https://doi.org/10.1177/20531680251356282