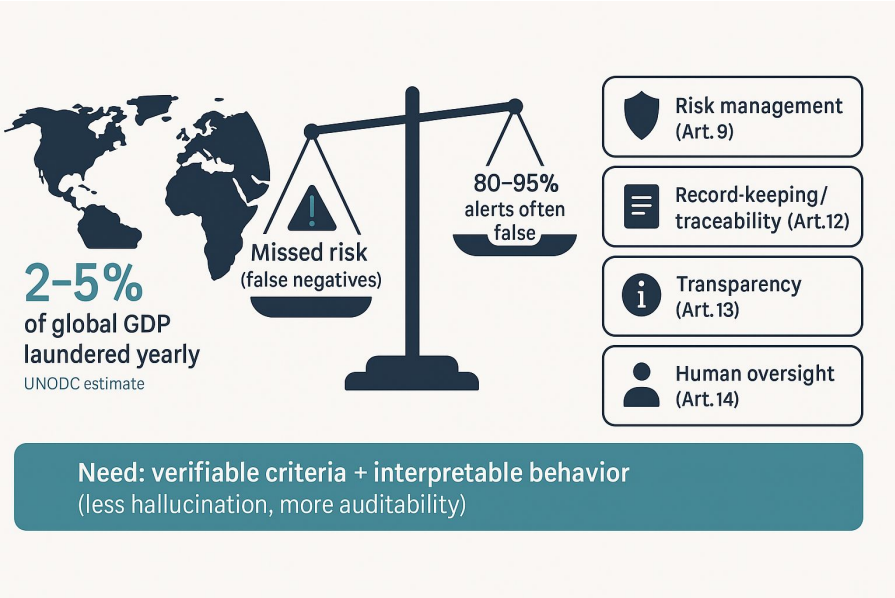
Research Methods and Professional Practice

Project Title: Interpretable Verifier-Reinforcement Learning for Financial Crime Detection via Sparse Autoencoders

Abdulhakim Bashir
October 2025

Supervisor: Dr Diego Navarra

2. Significance / Problem Context



The Financial Crime Detection Challenge

Scale: Global money laundering estimated at trillions annually (UNODC, 2023)

Cost: Compliance systems overwhelmed with false positives (FCA CP24/9, 2024)

Regulation: EU AI Act (2024) mandates transparency for high-risk AI

Gap: Black-box ML conflicts with auditability requirements (Casper et al., 2024)

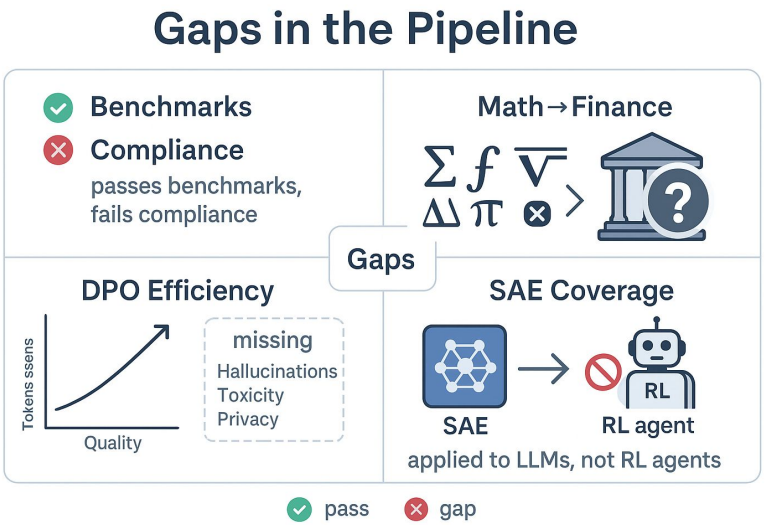
3. Research Gap

1. **Domain Evidence:** Alignment work is mostly tested on general chat/benchmarks; reviews call for domain-specific, auditable evidence (e.g., compliance/finance). **(Sharkey et al., 2025)**

2. Process supervision and verifier-based rewards have shown significant gains in math and code reasoning **(Lightman et al., 2023; Shao et al., 2024; DeepSeek AI, 2024)**, but the application of these early-stage techniques to regulated workflows remains an open question.

3. **Preference-only alignment dominates practice. DPO (Rafailov et al. in 2023)** is compute-light and widely adopted, but we lack studies on **calibration/abstention/citation** in high-stakes decision support. (Ji et al., 2024)

4. **Interpretability gap. Sparse autoencoders (SAEs)** recover monosemantic features in LLMs(Anthropic 2024), yet **little/no published analysis** looks at **SAE features on RL-aligned reasoning traces**.



4. Research Questions

Can Sparse Autoencoders reveal what reinforcement-learning with verifiable rewards actually teaches LLMs for financial-crime detection?

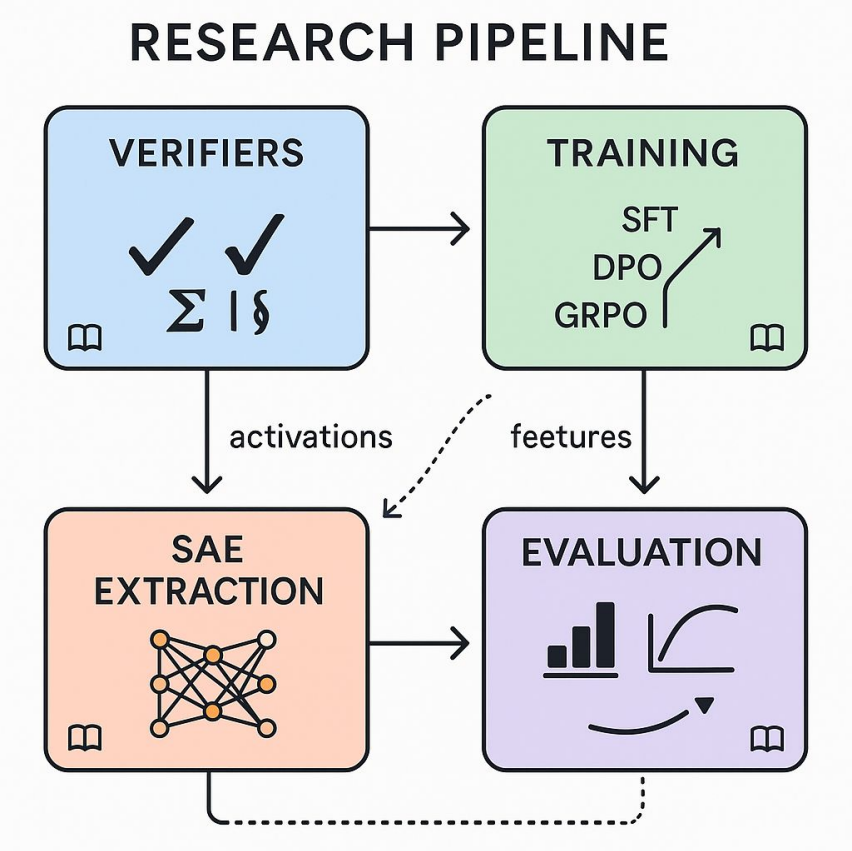
- **RQ1.** Does **verifiable-reward training** change internal representations vs **SFT** and **preference-only** alignment? (Motivated by process supervision results and RL-reasoning gains.)

- **RQ2.** Can **SAEs** surface **domain-relevant, monosemantic features** that correlate with **verifier success** (e.g., SAR elements, risk indicators, abstention)? (SAEs show monosemanticity at scale; early SAE-for-RL exists only on toy DQN(deep query network. DuPlessie, 2025)

- **RQ3.** What **label / compute** budget achieves the best **safety–helpfulness trade-off**? (Preference-only baselines like **DPO**/**SimPO** are compute-light and widely used.)
- **RQ4**: Can modular verifiable rewards enable regulatory adaptation without retraining?

5. Aims & Objectives

1. Build verifiers (Math [GSM8K – Cobbe et al., 2021]; AML rules – FinCEN guidance).
2. Train SFT → DPO (Rafailov et al., 2023) / SimPO (Meng et al., 2024) → GRPO (Shao et al., 2024).
3. Extract SAE features pre/post training (Anthropic, 2024).
4. Test causal influence and report efficiency + faithfulness metrics (e.g. Matton et al., 2025; Zaman & Srivastava, 2025; Chan et al., 2022).



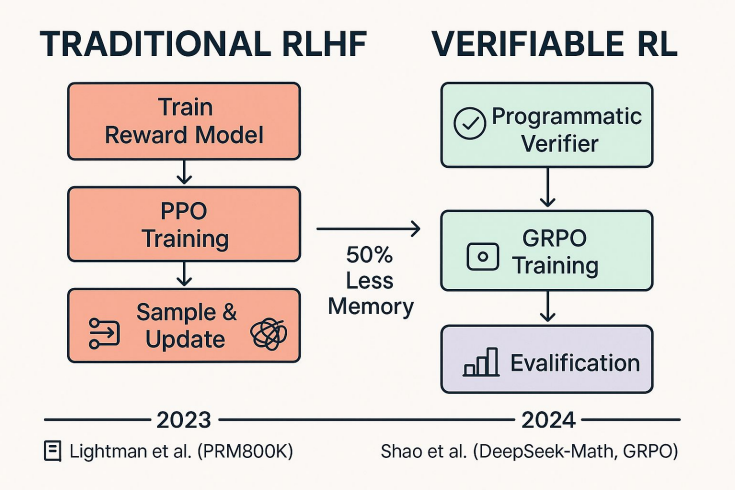
6. Key Literature

Three Research Streams

1. **Verifiable Rewards & RL:**
   - Process supervision (Lightman et al., 2023)
   - Programmatic correctness checking
   - Contrast: RLHF reward hacking vulnerabilities
2. **Preference Optimization:**
   - DPO: closed-form extraction (Rafailov et al., 2023)
   - Compute-efficient baseline
   - Safety properties: under-studied
3. **Sparse Autoencoders:**
   - Dictionary learning for interpretability
   - Anthropic's monosemanticity work (2023-2024)
   - Scaling to production models

6. Key Literature 1 - Verifiable reward & RL Reasoning

Process Supervision > Outcome Supervision
✓ OpenAI PRM800K: 78% on MATH (Lightman et al., 2023)
✓ Step-level feedback prevents reward hacking
Group Relative Policy Optimization (GRPO)
✓ DeepSeek-Math: 51.7% → 88.2% GSM8K (Shao et al., 2024)
✓ No critic model = 50% memory reduction
✓ Group-based advantage estimation

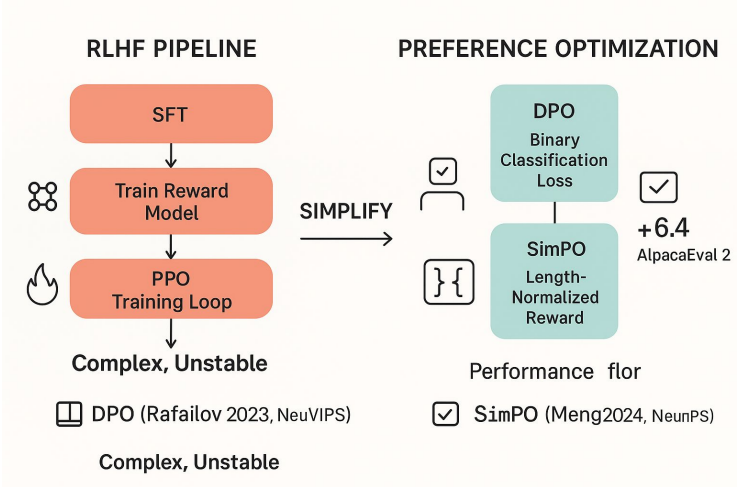


Key Insight: Verifiable signals (math, code, compliance) eliminate learned reward model vulnerabilities.

Application: AML rules = real-world verifiable criteria.

7. Key Literature 2 - Preference Optimization

Compute-Efficient Alternatives to RLHF
Direct Preference Optimization (DPO)
✓ Closed-form policy extraction (Rafailov et al., 2023)
✓ Single-stage training vs. two-stage RLHF
✓ Binary cross-entropy loss = simpler & stable
SimPO: Reference-Free Optimization
✓ Length-normalized reward (Meng et al., 2024)
✓ No reference model = 20% faster, 10% less memory
✓ +6.4 points over DPO on AlpacaEval 2 (Dubois et al., 2024)



RATIONALE: baseline comparison for verifiable rewards

7. Key Literature 3 - SAEs & Mechanistic Interpretability

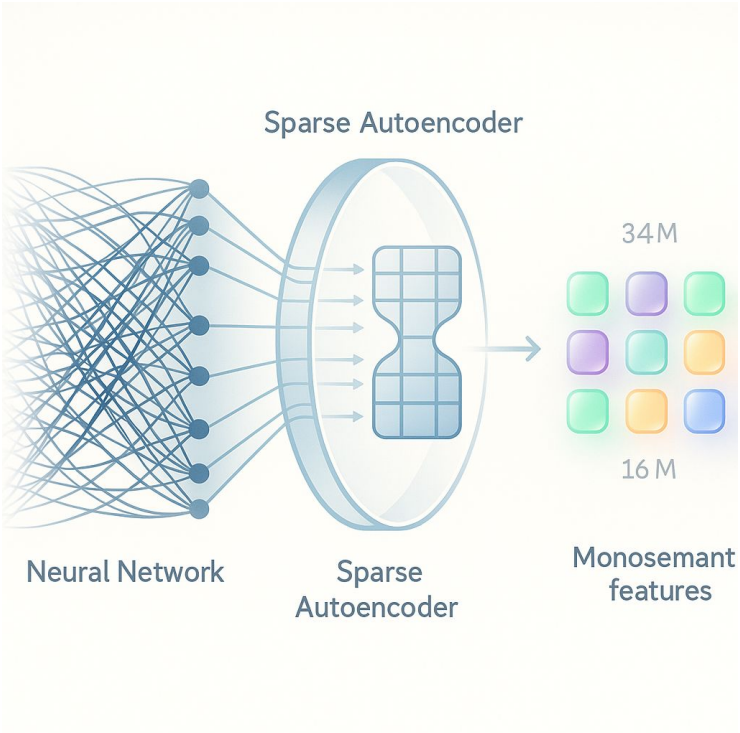
Solving Superposition via Dictionary Learning

- ✓ Anthropic: 34M monosemantic features (2024), (Rajamanoharan et al., 2024)
- ✓ OpenAI TopK SAEs: 16M latents, 7% dead (Gao 2024)
- ✓ Causal feature relevance via ablation tests

Application to RL Understanding

- ✓ Demircan et al. (2024): SAEs reveal TD-learning in LLMs
- ✓ Feature extraction → causal intervention → verification

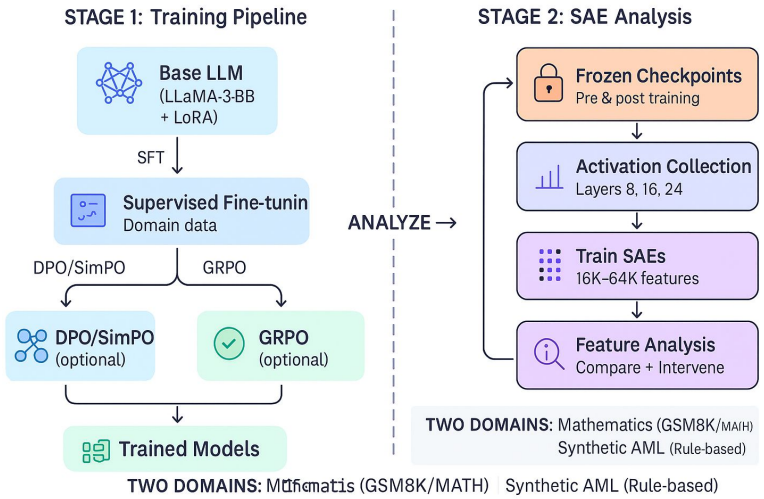
OBJECTIVE: Extract & analyze features from GRPO/DPO-trained models to understand what verification-driven training actually learns



7. Methodology Overview

METHODOLOGY: TWO-STAGE PIPELINE
STAGE 1: Training with Verifiable Rewards
Base Model: LLaMA-3-8B or Qwen-2.5-7B + LoRA (Hu et al. 2021)
├─ SFT: Domain-specific fine-tuning
├─ DPO/SimPO: Preference optimization baseline
└─ GRPO : Verifiable reward optimization

STAGE 2: SAE-Based Feature Analysis
├─ Freeze pre/post-training checkpoints
├─ Extract activations (layers 8, 16, 24)
├─ Train SAEs (16K-64K features per layer)
└─ Compare feature evolution + causal interventions
TWO DOMAINS:
- Mathematics (GSM8K/MATH) - exact verification
- Synthetic AML/Fin crime - rule-based compliance checks



8. Datasets & Verification

Mathematical Reasoning

- GSM8K dataset (Cobbe et al., 2021)
- Programmatic correctness checking

Financial Crime Detection

- IEEE-CIS Fraud Detection (590K transactions, Kaggle 2019)
- Synthetic SAR narratives (FinCEN guidance-based)

Verification Framework

- Rule-based compliance checkers
- Required narrative elements
- Abstention when evidence insufficient

Regulatory Adaptation

- Document retrieval for rule updates
- Modular reward structures
- Zero-shot adaptation capability

Privacy: 100% synthetic data; no real customer information

9. Training Plan

Parameter-Efficient Fine-Tuning

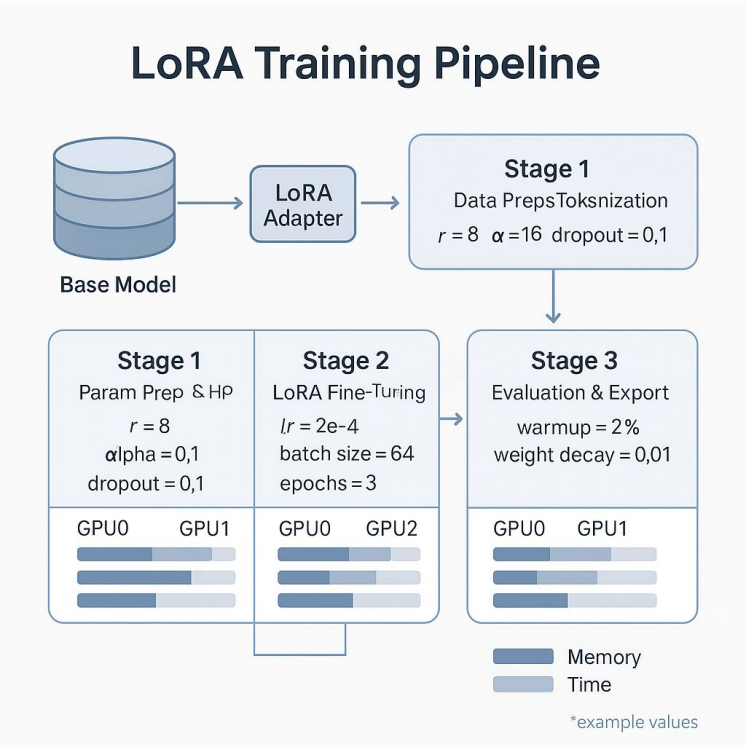
Base Models: Qwen-2.5-7B
LoRA: Rank 8-16; ~0.1% trainable parameters

Three Stages:

1. **SFT:** 3 epochs, lr=2e-5
2. **DPO:** β=0.1, 2 epochs, lr=1e-5
3. **RL :** Verifier-based rewards, 1-2 epochs
4. SAE on the internal activations to compare SFT vs DPO vs RL reasoning traces

Resources:

- 2× NVIDIA A100 GPUs (40GB)
- Mixed precision (bfloat16)
- 30-50 GPU-hours estimated



10. SAE Analysis Plan - Feature Extraction Strategy

Layer Selection:

- Layer 8: Early semantic processing
- Layer 16: Mid-network reasoning
- Layer 24: Late-stage decisions

SAE Configuration:

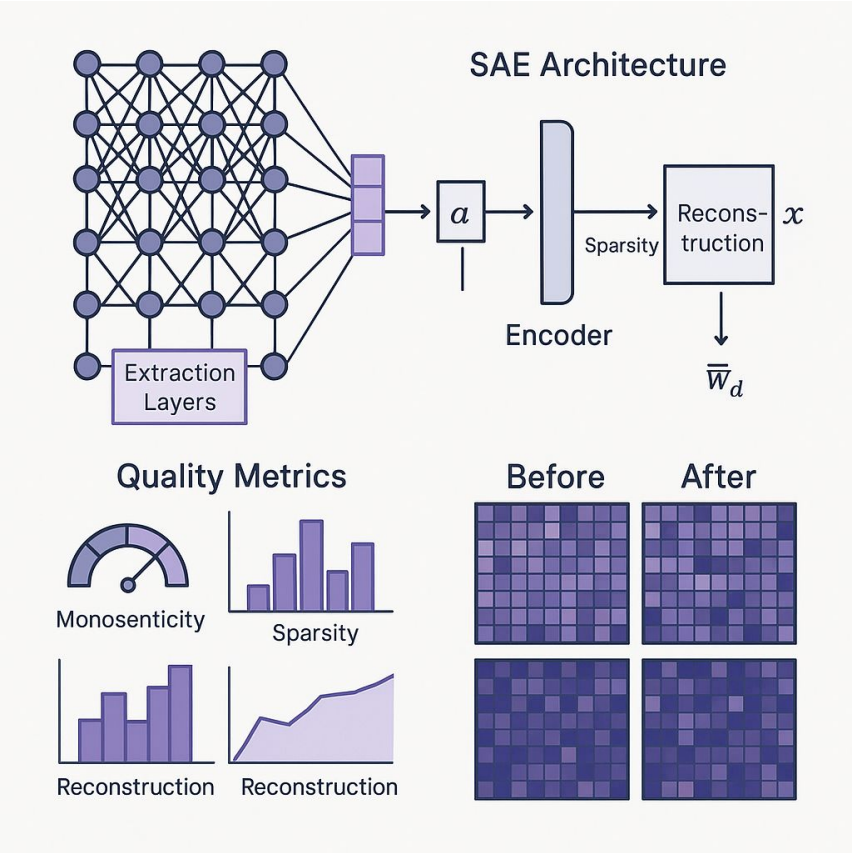
- 16K-64K features per layer (4×-16× overcomplete)
- L_1 sparsity penalty
- 100M activation samples

Quality Metrics:

- Monosemanticity scoring
- Sparsity: <50 features active/token
- Reconstruction loss <0.01 MSE
- Dead features <10%

Analysis:

- Pre vs. post-training comparison
- Causal ablation experiments



11. Evaluation & Metrics

Multi-Dimensional Assessment

Performance:

- Verifier pass rate (math: exact; AML: rule-based)
- Reasoning quality (chain-of-thought faithfulness)
- Abstention accuracy

Safety:

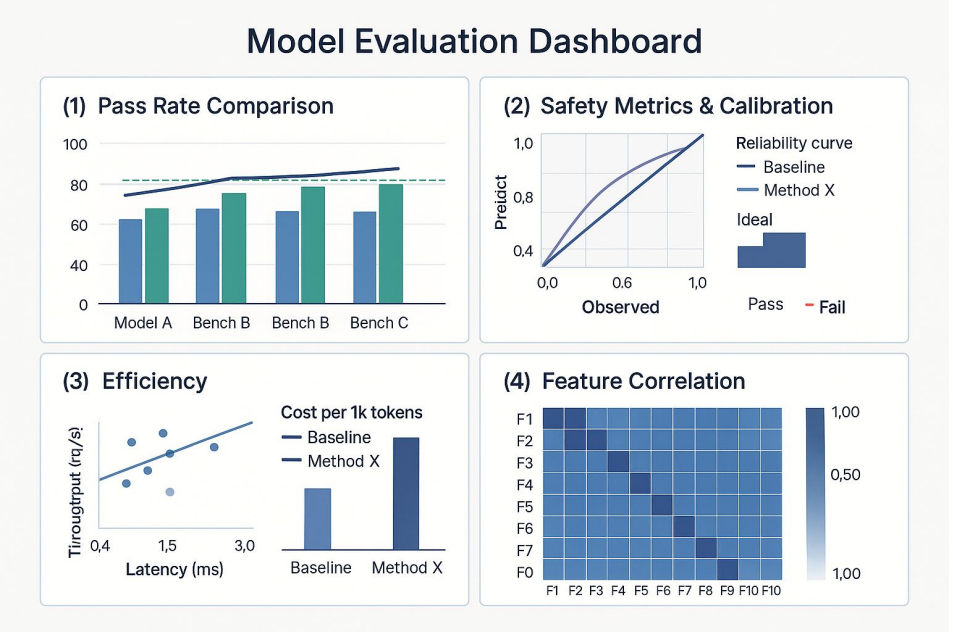
- Confidence calibration
- Out-of-distribution detection
- Hallucination rate

Efficiency:

- Label efficiency curves (pass rate vs. examples)
- Compute efficiency (pass rate vs. GPU-hours)
- SFT vs. DPO vs. RL comparison

Interpretability:

- Feature-verifier correlation
- Causal attribution via ablation



12. Ethical Considerations & Risks

Four Risk Categories

1. Privacy: 100% synthetic data; no real PII; GDPR-compliant (European Commission, 2016)

2. Bias:

- Fairness audits; protected attribute analysis;
- Demographic parity testing (Mitchell et al., 2021)

3. Misuse: No public deployment; controlled access; audit logging

4. Governance:

- Comprehensive logging (all experiments)
- Ethics approval (synthetic data category)
- Human oversight for high-stakes tasks (Floridi & Cowls, 2022)

13. Deliverable Artefacts

Open-Source Contributions

1. **Trained Models:** LoRA adapters (SFT, DPO, RL checkpoints)
2. **Verifiers:** Math correctness checker; AML rule engine
3. **Dataset:** 1,000+ synthetic SAR scenarios with labels
4. **SAE Toolkit:** Feature extraction; visualization; causal interventions
5. **Documentation:** Audit logs; ethics checklists; experimental protocols
6. **Academic:** MSc dissertation;
7. **Source code**: Used throughout.

License: Apache 2.0 with responsible use guidelines

15. Limitations

Technical

- Limited compute → small-scale RL & SAE
- GRPO instability → fallback to DPO (Rafailov et al., 2023)
- SAE ambiguity → correlation & case studies (Anthropic, 2024)

Methodological

- Synthetic AML data → limited realism
- Verifier ≠ full regulatory context (EU AI Act, 2024)
- Causal tests = indicative, not proof (Demircan et al., 2024)

Scope

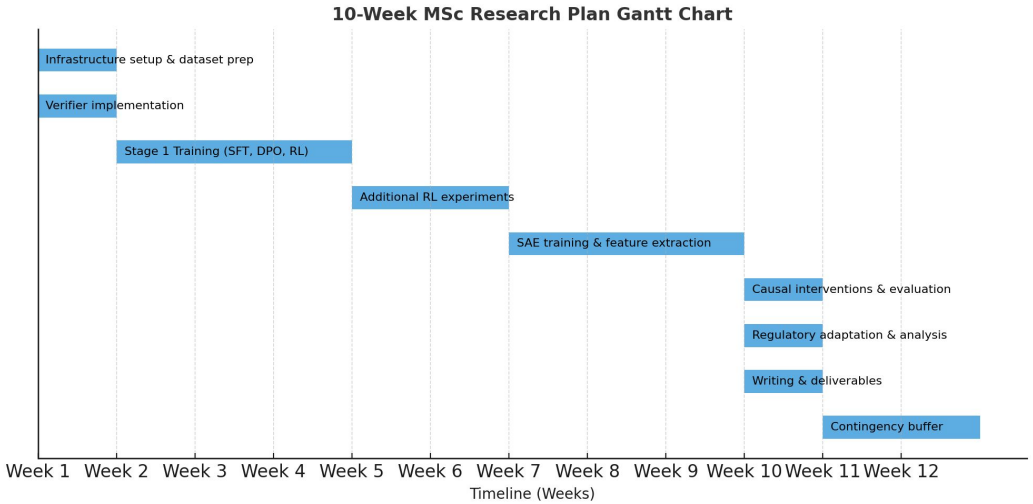
- Short MSc timeline → no deployment testing

14. Timeline

10-Week Research Plan

Weeks 1: Infrastructure setup; dataset preparation; verifier implementation
Weeks 2-4: Stage 1 training (SFT, DPO, optional RL)
Weeks 5-6: Additional RL experiments if needed
Weeks 7-9: SAE training and feature extraction
Week 10: Causal interventions and evaluation
Week 10: Regulatory adaptation testing, Analysis, writing & deliverables

Contingency: 2-week buffer for challenges



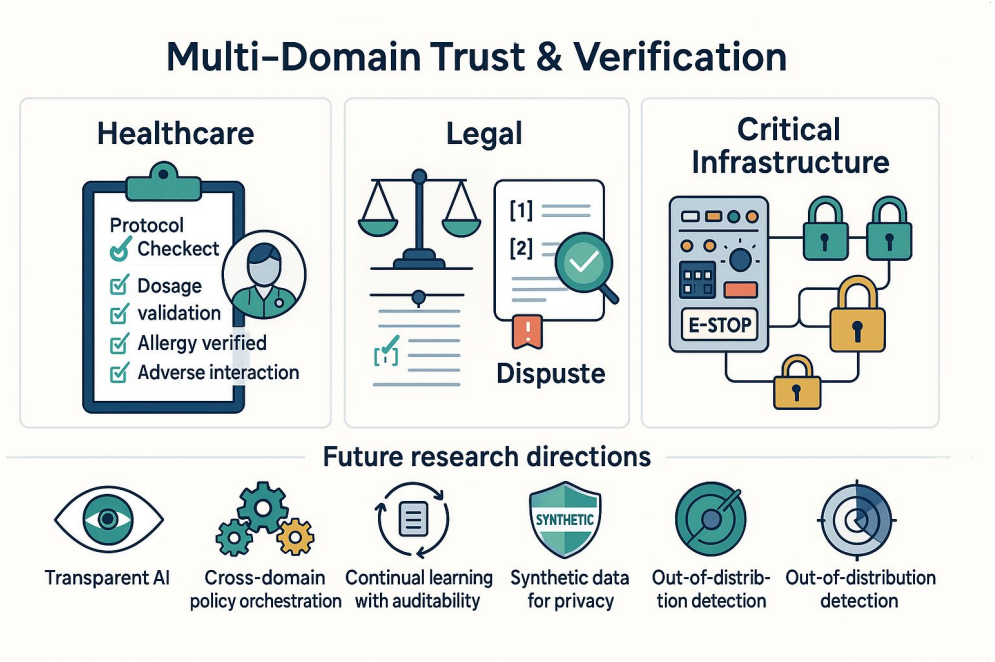
16. Broader Applicability

Cross-Domain Applications

Healthcare: Treatment protocol compliance; drug interactions
Legal: Precedent citation; contract compliance
Critical Infrastructure: Safety protocols; control sequences

Future Research:

- Multi-domain feature transfer
- Human-AI collaboration with SAE explanations
- Adversarial robustness analysis
- Scaling to 70B+ models
- Real-world deployment validation



17. Conclusion

Central Question: Can sparse autoencoders reveal what verifiable-reward RL teaches language models for high-stakes decision-making?

Approach:

- Domain validation (Financial crime, AML)
- Multi-stage training with verifiable rewards
- SAE-based mechanistic analysis
- Regulatory adaptation capability

Expected Impact:

- Interpretable AI for financial crime detection
- Reduced false positives, auditable decisions
- Framework for high-stakes AI deployment

Thank you.

References

Anthropic. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. https://transformer-circuits.pub/2023/monosemantic-features

Anthropic, 2024. *Scaling Monosemanticity: Extracting 34 Million Interpretable Features from Claude 3 Sonnet*. Anthropic Interpretability Research: Available at: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T.L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M. and et al. (2024) *Black-Box Access is Insufficient for Rigorous AI Audits*. Available at: https://dl.acm.org/doi/10.1145/3630106.3659037

Cobbe, K., Kosaraju, V., Bavarian, M., et al. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*. Available via https://arxiv.org/abs/2110.14168

Chan, K., Li, T., & Smith, J., 2022. *A Comparative Study of Faithfulness Metrics for Model Interpretability Methods*. In ACL 2022. Available at: https://aclanthology.org/2022.acl-long.345

DuPlessie, P. (2025) Sparse Autoencoders for Interpretability in Reinforcement Learning Models: Available at: https://math.mit.edu/research/highschool/primes/materials/2024/DuPlessie.pdf

Dubois, Y., Galambosi, B., Liang, P. and Hashimoto, T.B. (2024) *Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators*. Available at: https://arxiv.org/pdf/2404.04475

References

EU. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. *Official Journal of the European Union*.

European Commission (2016) *General Data Protection Regulation (GDPR)*. Available via: https://eur-lex.europa.eu/eli/reg/2016/679

FCA (FCA CP24/9, 2024) Available via: https://www.fca.org.uk/publication/consultation/cp24-9.pdf

FinCEN. (2020). *The Anti-Money Laundering Act of 2020*. U.S. Department of the Treasury. Available via : https://www.fincen.gov/anti-money-laundering-act-2020

Floridi, L. and Cowls, J. (2022) 'A unified framework of five principles for AI in society'. Available via: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3831321

Gao, L. et al. (2024), Scaling Sparse Autoencoders to 16 Million Features, OpenAI Technical Report, April 2024. Available at: https://arxiv.org/pdf/2406.04093

Hu, E. J., Shen, Y., Wallis, P., et al. (2021). LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

IEEE-CIS. (2019). *IEEE-CIS fraud detection*. Kaggle. https://www.kaggle.com/c/ieee-fraud-detection

Ji, M., Wu, Y., Wu, Z., Wang, S., Yang, J., Dras, M., & Naseem, U. (2024). *A Survey on Progress in LLM Alignment from the Perspective of Reward Design*. Available at: https://arxiv.org/pdf/2505.02666

References

Matton, L., Dubois, Y., & Bach, F., 2025. *Walk the Talk? Measuring the Faithfulness of Large Language Model Explanations*. Available at: https://arxiv.org/html/2504.14150

Mitchell, M. et al. (2021) 'Model cards for model reporting' Available via: https://www.semanticscholar.org/paper/Model-Cards-for-Model-Reporting-Mitchell-Wu/7365f887c938ca21a6adbef08b5a520ebbd4638f

Meng et al. (2024). *SimPO: Simple Preference Optimization with a Reference-Free Reward*. NeurIPS. Available at: https://papers.nips.cc/paper_files/paper/2024/file/e099c1c9699814af0be873a175361713-Paper-Conference.pdf

Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., Shah, R. and Nanda, N. (2024) Improving Dictionary Learning with Gated Sparse Autoencoders. Available at: https://arxiv.org/pdf/2404.16014

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J. and et al. (2025) *Open Problems in Mechanistic Interpretability*. Available at: https://arxiv.org/abs/2501.16496

Shao, Z., Wang, Y., et al. (2024) *DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models*. Available at: https://arxiv.org/pdf/2402.03300

Lightman, S., Kosaraju, V., et al. (2023) *Let's Verify Step by Step*. Available at: https://arxiv.org/abs/2305.20050.

References

Rafailov, R., Sharma, A., Mitchell, E., et al. (2023). Direct preference optimization: Your language model is secretly a reward model. *Proceedings of NeurIPS*.

UNODC. (2023). *Money-laundering and globalization*. United Nations Office on Drugs and Crime. https://www.unodc.org/unodc/en/money-laundering/

Zaman, H. & Srivastava, A., 2025. *A Causal Lens for Evaluating Faithfulness Metrics*. arXiv preprint Available at: https://arxiv.org/pdf/2502.18848