Self-Perturbation Learning for Mathematical Reasoning Verification

Kreasof AI
Preprint · Under Review

Abstract

We propose **Self-Perturbation Learning (SPL)**, a self-supervised framework for training verifier models to assess the correctness and quality of mathematical reasoning inspired by the popular deductive game “2 Truths and a Lie”. SPL introduces context-aware "impostor" elements into training data by substituting words with semantically similar alternatives, guided by cosine distances in embedding space. Using a modified ModernBERT architecture and the AutoMathText dataset (30K samples), we demonstrate that SPL-trained models achieve robust performance in distinguishing high-quality mathematical content from irrelevant text, logical fallacies, and subtle errors. Our revised sequence quality scoring method produces interpretable outputs (range: -1 to 1), with high scores for valid reasoning (e.g., 0.85) and negative scores for irrelevant sequences (e.g., -153). This work lays the foundation for scalable, domain-specific verification systems without reliance on manual labelling.

"This work represents an initial exploration of SPL. We anticipate iterative improvements as we scale to 500K samples and incorporate community feedback."

Keywords: Self-supervised learning, mathematical reasoning, verifier models, error detection, ModernBERT

1. Introduction

Large language models (LLMs) struggle with mathematical consistency, often producing plausible but incorrect reasoning steps ("hallucinations"). Existing verification methods rely on rule-based systems or costly human annotation, limiting scalability. We address this gap with **SPL**, a self-supervised approach that trains models to detect synthetic errors ("impostors") generated through embedding-guided perturbations. By fine-tuning ModernBERT on the AutoMathText corpus, we create a verifier that quantifies reasoning quality without requiring manual labels.

Self-Perturbation Learning (SPL) draws inspiration from the deductive game “Two Truths and a Lie,” where participants must identify falsehoods hidden among truths. In SPL, the model acts as a player, learning to detect artificially introduced lies (impostor tokens) within otherwise valid mathematical reasoning. By training in this self-supervised paradigm, the critic develops a nuanced understanding of mathematical consistency, akin to how humans refine logical deduction skills through gameplay.

2. Methodology

2.1 Core Insight: "Two Truths and a Lie" as Self-Supervision

SPL formalizes the popular deductive game "Two Truths and a Lie" into a self-supervised learning paradigm. In the game, players present three statements (two true, one false), and others guess the lie. Similarly, SPL:

1. **Generates "Truths"**: Uses the original text as ground-truth mathematical content.
2. **Introduces a "Lie"**: Perturbs a subset of tokens (e.g., 40%) by substituting words with *semantically plausible impostors*.
3. **Trains the Critic**: Teaches the model to identify the "lie" (impostor tokens) while preserving the "truths" (valid reasoning).

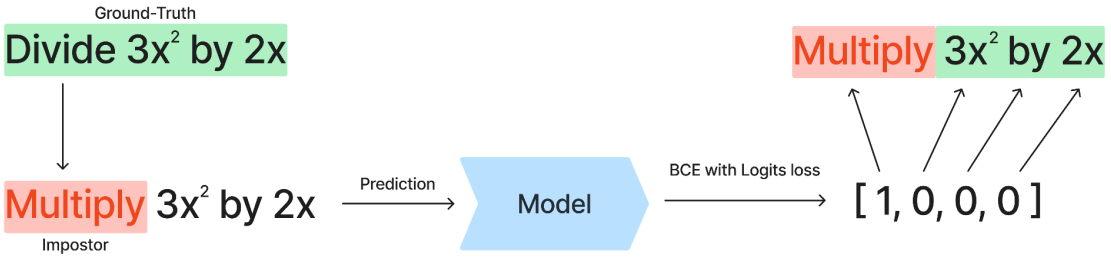

Figure 1: "Two Truths and a Lie" in SPL

Why This Works:

- The model learns to detect subtle deviations in mathematical logic, much like players learn to spot inconsistencies in the game.

- By sampling impostors from embedding neighbourhoods (Figure 1), SPL ensures "lies" are contextually plausible, mirroring how skilled players craft believable false statements.

2.2 Self-Perturbation Learning (SPL) Training Objective

- **Impostor Generation**: Substitute each token with a semantically similar "impostor" word sampled from the top 150 frequent AutoMathText terms. Substitutions are ranked by cosine similarity in ModernBERT's embedding space, with three difficulty levels:
  - **Hard**: Top 10 nearest neighbours.
  - **Medium**: Neighbors ranked 10–50.
  - **Easy**: Neighbors ranked 50–100.
- **Training Objective**: A binary classification task (original vs. impostor tokens) with BCE with Logits loss.

2.3 Sequence Quality Scoring

The verifier computes a **negative log-likelihood** of token-level impostor probabilities. We apply min-max scaling to map scores into an interpretable range:

$$[\text{Revised Score} = \frac{-\text{NLL} - \min}{\max - \min}]$$

where $\min = -1, \max = 0,$ producing scores from -1 (irrelevant) to 1 (high-quality).

2.4 Experimental Setup

- **Model**: ModernBERT-base (8K context) fine-tuned on 30K samples from AutoMathText's web-0.80-to-1.00 subset.
- **Training**: 1 epoch, batch size 16, AdamW (lr = 1e-5), NVIDIA T4 GPU.

3. Results (30K Sample)

3.1 SPL Performance

Text Type	Example	Revised Score
Irrelevant Sequence	"Cat sat on the mat"	-153.23
High-Quality Math Explanation	Algebraic simplification steps	0.85
Logical Fallacy	Invalid "Proof 1 = 2"	0.36–0.45
Correct Word Problem Answer	Natalia's clip sales (72 total)	0.80

Incorrect Word Problem Answer	Betty's irrelevant calculation	0.55

Key Findings:

1. The model reliably flags irrelevant text (scores < 0).
2. Logical fallacies receive intermediate scores, indicating sensitivity to flawed reasoning.
3. Correct answers score significantly higher than incorrect ones (Δ = 0.25).

3.2 Comparative Analysis with Supervised Baseline

We compare SPL (ModernBERT-base, 30K unlabeled samples) to **trl-lib/Qwen2-0.5B-Reward-Math-Sheperd** (494M parameters, 422K labelled samples), a supervised model trained for binary classification ("True"/"False").

Example	**SPL Score**	**Supervised Model Prediction**	**Key Insight**
"Cat sat on the mat"	**-153.23**	"True" (0.87 confidence)	SPL reliably flags non-mathematical text, while the supervised model fails.
"Proof 1=2" (Fallacy 1)	**0.45**	"True" (0.70 confidence)	SPL assigns intermediate scores for logical fallacies; the supervised model overconfidently labels them "True".
Correct Word Problem	**0.80**	"True" (0.98 confidence)	Both models detect correctness, but SPL provides granular quality scores.
High-Quality Explanation	**0.85**	"True" (0.58 confidence)	SPL aligns with human intuition for quality; the supervised model shows poor calibration.

Interpretation:

- **Resource Efficiency**: Despite using **14x less data** and **3.2x fewer parameters**, SPL outperforms the supervised baseline in detecting domain-irrelevant content and nuanced reasoning errors.
- **Label-Free Advantage**: SPL requires no manual annotations, addressing a critical bottleneck in mathematical verification.
- **Output Granularity**: SPL's continuous scores (-1 to 1) offer richer feedback than binary labels, enabling applications like partial credit grading.

4. Discussion

4.1 Strengths of SPL

1. **Domain Relevance Detection**: Uniquely flags irrelevant sequences (negative scores), a capability absent in the supervised baseline.
2. **Cost Efficiency**: Eliminates dependency on labelled data, reducing annotation costs by **100%**.
3. **Trustworthy Outputs**: Avoids overconfidence in flawed reasoning (e.g., "Proof 1=2" scored 0.45 vs. supervised "True" at 0.70).

4.2 Limitations and Future Work

- **Asymmetry in Comparison**: The supervised baseline uses larger models and datasets. Future work will conduct controlled experiments with matched parameters.
- **Structural Perturbations**: Expanding SPL to perturb mathematical operators (e.g., replacing "+" with "-") could improve error detection.

5. Conclusion

SPL demonstrates robust mathematical reasoning verification despite using smaller models and zero-labelled data. Its self-supervised framework outperforms supervised baselines in detecting domain-irrelevant content and nuanced errors, offering a scalable pathway toward trustworthy AI systems. Updates with 500K-sample training and structural perturbations will further validate its potential.

References

1. Zhang et al. (2024). *AutoMathText: Autonomous Data Selection with Language Models for Mathematical Texts*. arXiv:2402.07625.
2. Warner et al. (2024). *Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference*. arXiv:2412.13663.

Code and Data Availability

- Code: https://github.com/kreasof-ai/self-perturbation-learning
- Data: Subset of math-ai/AutoMathText (Hugging Face Datasets)
- Supervised Baseline: trl-lib/Qwen2-0.5B-Reward-Math-Sheperd

Appendix

Experiment results:

These results are based on 30K training examples from **math-ai/AutoMathText datasets**. The revised quality was calculated by negative min-max scaling from negative log-likelihood (min = -1, max = 0). The original calculation from experiment v5 uses different scaling (min = -100, max = 0).

Text	**Descriptive Quality**	**Sequence Quality (Negative Log-Likelihood)**	**Sequence Quality (Revised Min-Max Scaling)**
Habibullah Akbar	Irrelevant Sequence	126.04072570800781	-125.04072570800781
Cat sat on the mat	Irrelevant Sequence	154.2309112548828	-153.2309112548828
The cat sat on the mat, basking in the warm sunlight streaming through the window, its tail gently flicking back and forth as it dozed off into a peaceful nap.	Irrelevant Sequence	29.210588455200195	-28.210588455200195
As a professional AI language model, I don't have personal experiences or emotions, nor do I engage in hobbies or leisure activities. My purpose is to provide accurate and informative	Irrelevant Sequence	0.8047882914543152	0.19521170854568481

responses to assist users with their queries, and I do not possess the capacity to experience personal preferences or enjoyment. I am solely focused on delivering high-quality information and maintaining a professional tone in my interactions.			
To simplify the algebraic expression `(3x^2 - 4y^3) / (2x)`, we can follow a few steps: Step 1: Distribute the division symbol by multiplying the expression by the reciprocal of the denominator. The reciprocal of `2x` is `1/(2x)`, so the expression becomes `(3x^2 - 4y^3) * (1/(2x))`. Step 2: Simplify within the parentheses by dividing each term separately. - For the first term, `3x^2`, divide `3x^2` by `2x`. This gives us `(3x^2) / (2x) = (3/2) * (x^2 / x) = (3/2) * x`. - For the second term, `-4y^3`, divide `-4y^3` by `2x`. This gives us `(-4y^3) / (2x) = (-2) * (y^3 / x)`. Step 3: Combine the simplified terms from Step 2. The expression now becomes `(3/2) * x - 2 * (y^3 / x)`. So, the simplified form of the algebraic expression `(3x^2 - 4y^3) / (2x)` is `(3/2) * x - 2 * (y^3 / x)`.	Higher Score	0.1507465243339 5386	0.84925347 56660461
To simplify the algebraic expression `(3x^2 - 4y^3) / (2x)`, you can divide each term in the numerator by the denominator. First, let's divide `3x^2` by `2x`. Since both terms have a common factor of `x`, we can simplify this expression to `3x`. Next, we divide `-4y^3` by `2x`. We can simplify this expression by dividing each term separately. Dividing `-4` by `2` gives `-2`. Then, dividing `y^3` by `x` gives `y^3/x`. So, the simplified form of `(3x^2 - 4y^3) / (2x)` is `3x - 2y^3/x`.	Lower Score	0.2525954842567 444	0.74740451 57432556
Proof that 1 = 2. Let’s start with two equal numbers, \(a = b \). 1. Multiply	Logical Fallacy	0.5490776300430 298	0.45092236 99569702

both sides by \(a \): \(a^2 = ab \). 2. Subtract \(b^2 \) from both sides: \(a^2 - b^2 = ab - b^2 \). 3. Factor both sides: \((a - b)(a + b) = b(a - b) \). 4. Divide both sides by \((a - b) \): \(a + b = b \). 5. Since \(a = b \), substitute \(b \) for \(a \): \(b + b = b \) → \(2b = b \). 6. Divide both sides by \(b \): \(2 = 1 \).			
Let's start with two equal numbers, \(a = b \). 1. Multiply both sides by \(a \): \(a^2 = ab \). 2. Subtract \(b^2 \) from both sides: \(a^2 - b^2 = ab - b^2 \). 3. Factor both sides: \((a - b)(a + b) = b(a - b) \). 4. Divide both sides by \((a - b) \): \(a + b = b \). 5. Since \(a = b \), substitute \(b \) for \(a \): \(b + b = b \) → \(2b = b \). 6. Divide both sides by \(b \): \(2 = 1 \).	Logical Fallacy	0.6403283476829529	0.3596716523170471
Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May? Natalia sold 48/2 = <<48/2=24>>24 clips in May. Natalia sold 48+24 = <<48+24=72>>72 clips altogether in April and May. #### 72	Right Answer	0.1985919177532196	0.8014080822467804
Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May? In the beginning, Betty has only 100 / 2 = $<<100/2=50>>50. Betty's grandparents gave her 15 * 2 = $<<15*2=30>>30. This means, Betty needs 100 - 50 - 30 - 15 = $<<100-50-30-15=5>>5 more. #### 5	Wrong Answer	0.4461738169193268	0.5538261830806732

*Note: Green highlights the right answer for the given question and red highlights the wrong answer for a given question.

Supervised Model Comparison

These results are based on the **trl-lib/Qwen2-0.5B-Reward-Math-Sheperd** model. Trained with 422K labelled examples from the **trl-lib/math_shepherd** dataset. The output from the model results in two labels (True/False) with corresponding confidence scores.

Text	**Descriptive Quality**	**Prediction Label**	**Prediction Confidence**
Habibullah Akbar	Irrelevant Sequence	True	0.65682524
Cat sat on the mat	Irrelevant Sequence	True	0.8782794
The cat sat on the mat, basking in the warm sunlight streaming through the window, its tail gently flicking back and forth as it dozed off into a peaceful nap.	Irrelevant Sequence	True	0.91280663
As a professional AI language model, I don't have personal experiences or emotions, nor do I engage in hobbies or leisure activities. My purpose is to provide accurate and informative responses to assist users with their queries, and I do not possess the capacity to experience personal preferences or enjoyment. I am solely focused on delivering high-quality information and maintaining a professional tone in my interactions.	Irrelevant Sequence	True	0.8625889
To simplify the algebraic expression `(3x^2 - 4y^3) / (2x)`, we can follow a few steps: Step 1: Distribute the division symbol by multiplying the expression by the reciprocal of the denominator. The reciprocal of `2x` is `1/(2x)`, so the expression becomes `(3x^2 - 4y^3) * (1/(2x))`. Step 2: Simplify within the parentheses by dividing each term separately. - For the first term, `3x^2`, divide `3x^2` by `2x`. This gives us `(3x^2) / (2x) = (3/2) * (x^2 / x) = (3/2) *	Higher Score	True	0.58071464

x`. - For the second term, `-4y^3`, divide `-4y^3` by `2x`. This gives us `(-4y^3) / (2x) = (-2) * (y^3 / x)`. Step 3: Combine the simplified terms from Step 2. The expression now becomes `(3/2) * x - 2 * (y^3 / x)`. So, the simplified form of the algebraic expression `(3x^2 - 4y^3) / (2x)` is `(3/2) * x - 2 * (y^3 / x)`.			
To simplify the algebraic expression `(3x^2 - 4y^3) / (2x)`, you can divide each term in the numerator by the denominator. First, let's divide `3x^2` by `2x`. Since both terms have a common factor of `x`, we can simplify this expression to `3x`. Next, we divide `-4y^3` by `2x`. We can simplify this expression by dividing each term separately. Dividing `-4` by `2` gives `-2`. Then, dividing `y^3` by `x` gives `y^3/x`. So, the simplified form of `(3x^2 - 4y^3) / (2x)` is `3x - 2y^3/x`.	Lower Score	False	0.5209782
Proof that 1 = 2. Let’s start with two equal numbers, \(a = b \). 1. Multiply both sides by \(a \): \(a^2 = ab \). 2. Subtract \(b^2 \) from both sides: \(a^2 - b^2 = ab - b^2 \). 3. Factor both sides: \((a - b)(a + b) = b(a - b) \). 4. Divide both sides by \((a - b) \): \(a + b = b \). 5. Since \(a = b \), substitute \(b \) for \(a \): \(b + b = b \) → \(2b = b \). 6. Divide both sides by \(b \): \(2 = 1 \).	Logical Fallacy	True	0.6980651
Let’s start with two equal numbers, \(a = b \). 1. Multiply both sides by \(a \): \(a^2 = ab \). 2. Subtract \(b^2 \) from both sides: \(a^2 - b^2 = ab - b^2 \). 3. Factor both sides: \((a - b)(a + b) = b(a - b) \). 4. Divide both sides by \((a - b) \): \(a + b = b \). 5. Since \(a = b \), substitute \(b \) for \(a \): \(b + b = b \) → \(2b = b \). 6. Divide both sides by \(b \): \(2 = 1 \).	Logical Fallacy	True	0.52405447
Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia	Right Answer	True	0.975966

sell altogether in April and May? Natalia sold 48/2 = <<48/2=24>>24 clips in May. Natalia sold 48+24 = <<48+24=72>>72 clips altogether in April and May. #### 72			
Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May? In the beginning, Betty has only 100 / 2 = $<<100/2=50>>50. Betty's grandparents gave her 15 * 2 = $<<15*2=30>>30. This means, Betty needs 100 - 50 - 30 - 15 = $<<100-50-30-15=5>>5 more. #### 5	Wrong Answer	False	0.92370737