Process vs. Outcome Supervision for Trustworthy and Interpretable LLM Reasoning: Research Resources

Overview

This comprehensive research compilation identifies 70+ papers and 25+ GitHub repositories across seven key research areas supporting work on process versus outcome supervision for trustworthy and interpretable LLM reasoning. Process supervision—providing feedback on each intermediate reasoning step rather than only final answers—has emerged as a critical technique for improving both accuracy and alignment in complex reasoning tasks, ACL Anthology +8 with implications spanning technical AI research, interpretability, and regulatory compliance.

---

1. Process Supervision Research

Foundational Papers

Solving math word problems with process- and outcome-based feedback
Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, Irina Higgins (2022)
https://arxiv.org/abs/2211.14275

Relevance: First comprehensive comparison between process-based and outcome-based supervision for language models on natural language reasoning tasks using GSM8K. Semantic Scholar Substack While outcome supervision achieves similar final-answer accuracy with less labeling, process supervision is necessary for ensuring correct reasoning steps. arXiv +2

Key Findings:

- Improved results from 16.8% → 12.7% final-answer error and 14.0% → 3.4% reasoning error among correct solutions arXiv Semantic Scholar
- Demonstrated that learned reward models can emulate process-based feedback effectively
- Established that detecting reasoning errors requires step-level supervision

---

Let's Verify Step by Step
Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, Karl Cobbe (2023)
https://arxiv.org/abs/2305.20050

Relevance: OpenAI's foundational work demonstrating that process supervision significantly outperforms outcome supervision for training reliable models on complex reasoning, achieving state-of-the-art on the MATH dataset. OpenAI Released PRM800K, the first large-scale process supervision dataset. OpenAI +4

Key Findings:

- Process-supervised models solve 78% of MATH problems vs. 72% for outcome supervision OpenAI +4
- Released 800,000 step-level human feedback labels (PRM800K dataset) arXiv +6
- Active learning significantly improves efficacy of process supervision arXiv +4
- Process supervision provides alignment benefits by directly rewarding human-endorsed reasoning chains OpenAI

Process-based Self-Rewarding Language Models

Shimao Zhang, Xiao Liu, Xin Zhang, Junxiao Liu, Zheheng Luo, Shujian Huang, Yeyun Gong (2025)
https://aclanthology.org/2025.findings-acl.930/ | https://arxiv.org/abs/2503.03746

Relevance: Addresses limitations of existing self-rewarding paradigms in mathematical reasoning by introducing long-thought reasoning, step-wise LLM-as-a-Judge, and step-wise preference optimization. Demonstrates that process-based self-rewarding can achieve reasoning capabilities surpassing human performance without human-annotated preference data. ACL Anthology arXiv

Key Findings:

- Successfully enhances LLM performance on multiple math benchmarks through iterative Process-based Self-Rewarding ACL Anthology arXiv
- Introduces step-wise evaluation and preference optimization within self-rewarding paradigm ACL Anthology
- Overcomes bottleneck of human performance limits in training data

---

Automated Process Supervision

Improve Mathematical Reasoning in Language Models by Automated Process Supervision (OmegaPRM)

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, Abhinav Rastogi (2024)
https://arxiv.org/abs/2406.06592

Relevance: Addresses scalability by proposing OmegaPRM, a divide-and-conquer MCTS algorithm that automates high-quality process supervision data collection without expensive human annotation or per-step Monte Carlo estimation. arXiv arXiv

Key Findings:

- Collected over 1.5 million process supervision annotations automatically—largest process supervision dataset (arXiv)
- Improved Gemini Pro from 51% to 69.4% on MATH500 and 86.4% to 93.6% on GSM8K (arXiv)
- Fully automated process operates without human intervention, making it cost-effective (arXiv)

Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations
Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, Zhifang Sui (2024)
https://aclanthology.org/2024.acl-long.510/

Relevance: Presents process-oriented math process reward model using automatically constructed process-wise supervision data, breaking the bottleneck of manual annotation. (Substack) Effective for both verification (reranking outputs) and reinforcement learning. (ACL Anthology) (arXiv)

Key Findings:

- Process RL with Math-Shepherd enhances Mistral-7B from 77.9% to 84.1% on GSM8K and 28.6% to 33.0% on MATH (arXiv)
- Accuracy improves to 89.1% and 43.5% with verification
- Demonstrates significant potential for automatic process supervision

---

Advanced Process Reward Models

Rewarding Progress: Scaling Automated Process Verifiers for LLM Reasoning (Process Advantage Verifiers)
Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Alekh Agarwal, Jonathan Berant (2024)
https://arxiv.org/abs/2410.08146

Relevance: Introduces novel approach measuring "progress"—the change in likelihood of producing correct response before and after taking a step. Train Process Advantage Verifiers (PAVs) showing that even weaker prover policies can substantially improve stronger base policies. (ADS) (Cool Papers)

Key Findings:

- Test-time search against PAVs is >8% more accurate and 1.5-5× more compute-efficient vs. Outcome Reward Models (arXiv +2)
- Online RL with dense rewards from PAVs achieves 5-6× gain in sample efficiency and >6% gain in accuracy (arXiv +2)
- Theoretical demonstration that weak prover policies can improve stronger base policies (OpenReview)

Process Reward Models That Think (ThinkPRM)

Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, Lu Wang (2025)
https://arxiv.org/abs/2504.16828

Relevance: Proposes data-efficient generative process reward model that verifies solutions step-by-step using verification chain-of-thought. Leverages reasoning abilities of long CoT models and is fine-tuned on orders of magnitude fewer process labels. (arXiv +5)

Key Findings:

- Outperforms discriminative PRMs trained on 100× more data using only 1% of process labels from PRM800K (8K labels) (Hugging Face +6)
- Surpasses discriminative verifiers trained on full PRM800K by 8% on GPQA-Diamond and 4.5% on LiveCodeBench (Hugging Face +4)
- Enables scaling of verification compute both in parallel and sequentially

---

The Lessons of Developing Process Reward Models in Mathematical Reasoning

Zhenru Zhang et al. (2025)
https://arxiv.org/abs/2501.07301

Relevance: Comprehensive study examining challenges and best practices in developing effective PRMs. Demonstrates that Monte Carlo estimation-based data synthesis typically yields inferior performance compared to LLM-as-a-judge and human annotation methods. (Hugging Face +3)

Key Findings:

- MC estimation-based data synthesis yields inferior performance and generalization vs. LLM-as-a-judge and human annotation (arXiv) (arXiv)
- Identifies three key biases in Best-of-N evaluation strategies (arXiv)
- Develops consensus filtering mechanism integrating MC estimation with LLM-as-a-judge (arXiv) (arXiv)

Enhancing Reasoning through Process Supervision with Monte Carlo Tree Search

Shuangtao Li, Xinhan Di, Wangzhi Deng, Zhijian Zhao (2025)
https://arxiv.org/abs/2501.01478

Relevance: Explores using MCTS to generate process supervision data with LLMs themselves for training. Samples reasoning steps and assigns scores capturing "relative correctness" through iterative generate-then-train cycles. (arXiv) (arXiv)

Key Findings:

- MCTS automatically generates high-quality process supervision labels without human annotation
- Shows considerable improvements on GSM8K and MATH (arXiv)
- Models trained on one dataset exhibit improved performance on the other, showing transferability (arXiv)

Foundational Outcome Supervision (Baseline)

Training Verifiers to Solve Math Word Problems

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, John Schulman (2021)
https://arxiv.org/abs/2110.14168

Relevance: Foundational paper from OpenAI introducing concept of training verifier models (outcome reward models) to rank multiple candidate solutions. (arxiv) (github) Established outcome reward models (ORMs) as baseline for comparison with process supervision methods. (Substack)

Key Findings:

- First systematic study of using trained verifiers in mathematical reasoning
- Demonstrated verifiers can rank solutions more effectively than human evaluators in some cases
- Established framework for using learned reward models to improve reasoning performance

---

2. Interpretability and Faithfulness

Chain-of-Thought Faithfulness

Measuring Faithfulness in Chain-of-Thought Reasoning

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, et al. (2023)
https://arxiv.org/abs/2307.13702

Relevance: Seminal paper investigating whether Chain-of-Thought reasoning faithfully reflects the model's actual reasoning process. (AI Alignment Forum) (arxiv) Addresses core question of whether stated reasoning steps genuinely explain how LLMs arrive at answers.

Key Findings:

- Introduces intervention-based methods to measure faithfulness by adding mistakes or paraphrasing CoT steps
- Models show large variation across tasks in how strongly they condition on CoT

- Reveals inverse scaling: larger, more capable models produce less faithful reasoning on most tasks

---

Faithful Chain-of-Thought Reasoning
Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Bhaskar Ramasubramanian, Arman Cohan (2023)
https://arxiv.org/abs/2301.13379

Relevance: Proposes novel framework guaranteeing faithfulness by design through symbolic reasoning, addressing the problem that standard CoT may generate plausible but unfaithful explanations. arXiv

Key Findings:

- Two-stage Faithful CoT: Translation (NL → symbolic reasoning chain) and Problem Solving (reasoning chain → answer)
- Achieves 6.3% accuracy gain on Math Word Problems, 3.4% on Planning, 5.5% on Multi-hop QA, and 21.4% on Relational Inference
- Sets new state-of-the-art few-shot performance on 7 datasets with 95.0+ accuracy on 6

Making Reasoning Matter: Measuring and Improving Faithfulness of Chain-of-Thought Reasoning
Debjit Paul, Robert West, Antoine Bosselut, Boi Faltings (2024)
https://arxiv.org/abs/2402.13950 | https://aclanthology.org/2024.findings-emnlp.882/

Relevance: Introduces causal mediation analysis to rigorously measure how LLM reasoning steps influence final answers, providing principled framework for quantifying faithfulness. ACL Anthology +2

Key Findings:

- Causal mediation analysis across 12 LLMs finds models do not reliably use intermediate reasoning steps
- Introduces FRODO framework with inference module (generates correct reasoning) and reasoning module (faithfully reasons over steps) ACL Anthology
- Instruction-tuned models show better faithfulness; vanilla LMs (<20B) systematically unfaithful

---

On Measuring Faithfulness or Self-consistency of Natural Language Explanations
Letitia Parcalabescu, Anette Frank (2024)
https://arxiv.org/abs/2311.07466

Relevance: Provides critical conceptual clarification by distinguishing faithfulness (alignment with model internals) from self-consistency (output-level coherence), arguing existing tests measure the latter not the former. arXiv arXiv

- Existing faithfulness tests measure self-consistency at output level rather than true faithfulness
- Constructs Comparative Consistency Bank comparing 11 open LLMs and 5 tasks across existing self-consistency tests
- Introduces CC-SHAP, a fine-grained measure comparing how model's input contributes to both predicted answer and explanation generation

---

Chain-of-Thought Reasoning In The Wild Is Not Always Faithful

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, Arthur Conmy (2025)
https://arxiv.org/abs/2503.08679

Relevance: Recent work demonstrating that unfaithful CoT occurs even without artificial bias in prompts, revealing implicit post-hoc rationalization in production models. arXiv arXiv

Key Findings:

- Models sometimes produce coherent arguments to justify systematically contradictory answers
- Labels phenomenon as "Implicit Post-Hoc Rationalization" due to implicit biases
- Measures surprisingly high rates in production models: GPT-4o-mini (13%), Haiku 3.5 (7%), Gemini 2.5 Flash (2.17%)

Measuring Chain of Thought Faithfulness by Unlearning Reasoning Steps

Martin Tutek, Saeed Amizadeh, Kamilė Lukošiūtė, Ivan Vulić (2025)
https://arxiv.org/abs/2502.14829

Relevance: Introduces novel parametric faithfulness framework using unlearning to measure whether CoT steps are truly influential, addressing limitation that contextual perturbations don't remove knowledge from parameters.

Key Findings:

- Faithfulness by Unlearning Reasoning steps (FUR) erases information from model parameters
- Demonstrates FUR can precisely change underlying models' predictions by unlearning key steps
- More accurately identifies when CoT is parametrically faithful compared to contextual intervention methods

---

Towards Better Chain-of-Thought: A Reflection on Effectiveness and Faithfulness
Jiachun Li, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, Jun Zhao (2024)
https://arxiv.org/abs/2405.18915

Relevance: Comprehensive analysis of CoT from dual perspectives of effectiveness (performance improvement) and faithfulness (alignment with reasoning), identifying key factors influencing both dimensions. arXiv arxiv

Key Findings:

- Identifies key factors: problem difficulty, information gain, and information flow
- Interprets unfaithful CoT through joint analysis of information interaction among question, CoT, and answer
- Proposes novel algorithm that recalls extra information from questions to enhance CoT generation

---

On the Hardness of Faithful Chain-of-Thought Reasoning in Large Language Models
Sree Harsha Tanneru, Chirag Agarwal, Himabindu Lakkaraju (2024)
https://arxiv.org/abs/2406.10625

Relevance: Rigorously investigates fundamental difficulty of eliciting faithful CoT, testing three major approaches (in-context learning, fine-tuning, activation editing) and revealing current methods are insufficient.

Key Findings:

- Tests novel strategies for in-context learning, fine-tuning, and activation editing
- Strategies offer only limited success with slight performance enhancements in controlled scenarios
- Underscores inherent difficulty of faithful CoT across all tested approaches

---

Step-Wise Formal Verification for LLM-Based Mathematical Problem Solving
Kuo Zhou, Yiheng He (2025)
https://arxiv.org/abs/2505.20869

Relevance: Introduces formal verification methods using symbolic tools (CAS, SMT solver) to evaluate correctness of each reasoning step, providing rigorous auditability. arXiv The Moonlight

Key Findings:

- Proposes MATH-VF framework with Formalizer (translates NL solution to formal context) and Critic (evaluates correctness using external tools) arXiv The Moonlight
- Demonstrates superior performance on MATH500 and ProcessBench benchmarks

- Provides corrective feedback when errors detected, enabling step-by-step formal verification

---

Evaluating Readability and Faithfulness of Concept-based Explanations

Meng Li, Haoran Jin, Ruixuan Huang, Zhihao Xu, Defu Lian, Zijia Lin, Di Zhang, Xiting Wang (2024)
📄 https://aclanthology.org/2024.emnlp-main.36/

Relevance: Addresses how to properly evaluate explanations by introducing formal definitions and metrics for both readability and faithfulness of concept-based explanations for LLMs.

Key Findings:

- Introduces formal definition of concepts generalizing to diverse explanation settings ACL Anthology
- Quantifies faithfulness via perturbation with optimization-based adequate perturbation in high-dimensional space ACL Anthology
- Applies meta-evaluation method from measurement theory generalizable to other explanation types

---

3. Mathematical and Multi-Step Reasoning

Core Datasets

GSM8K: Grade School Math 8K

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. (2021)
📄 https://arxiv.org/abs/2110.14168
🔗 GitHub: https://github.com/openai/grade-school-math
🤗 HuggingFace: https://huggingface.co/datasets/openai/gsm8k

Description: 8.5K high-quality, linguistically diverse grade school math word problems arXiv (7.5K train, 1K test). arXiv Problems require 2-8 steps using basic arithmetic. arxiv github Solutions provided in natural language with step-by-step reasoning chains and calculator annotations. Hugging Face +2

Key Characteristics:

- Gold standard for evaluating basic mathematical reasoning in LLMs GitHub
- Natural language solutions with explicit reasoning chains
- 350,000+ monthly downloads as of Feb 2025 gradient science
- Problems written by human contractors with quality control github +2

MATH: Mathematics Aptitude Test of Heuristics

Dan Hendrycks, Collin Burns, Saurav Kadavath, et al. (2021)

📄 https://arxiv.org/abs/2103.03874
🔗 GitHub: https://github.com/hendrycks/math
🤗 HuggingFace: https://huggingface.co/datasets/hendrycks/competition_math

Description: 12,500 challenging competition-level mathematics problems (7,500 train, 5,000 test) spanning seven subjects: Prealgebra, Algebra, Number Theory, Counting and Probability, Geometry, Intermediate Algebra, and Precalculus. arXiv Sourced from AMC 10, AMC 12, and AIME competitions. arXiv +2

Key Characteristics:

- Significantly more challenging than GSM8K; tests advanced mathematical reasoning GitHub
- Full step-by-step solutions in LaTeX arXiv Hugging Face
- Difficulty levels 1-5 allow granular performance analysis
- Includes AMPS pretraining dataset

PRM800K: Process Reward Model Dataset
Hunter Lightman et al. (2023)
📄 https://arxiv.org/abs/2305.20050
🔗 GitHub: https://github.com/openai/prm800k
🤗 HuggingFace: https://huggingface.co/datasets/openai/prm800k

Description: 800,000 step-level correctness labels for model-generated solutions to MATH problems. GitHub Human labelers annotated each intermediate reasoning step with positive, negative, or neutral labels. GitHub +4

Key Characteristics:

- Enables training of step-by-step verifiers for mathematical reasoning Substack
- Human-verified labels at granularity of individual reasoning steps PyPI
- Active learning strategy for data collection efficiency Gabormelli +2
- Used extensively in recent work on test-time scaling and verification

---

Benchmarks and Evaluation

Large Language Models for Mathematical Reasoning: Progresses and Challenges (Survey)
Jaehun Ahn et al. (2024)
📄 https://arxiv.org/abs/2402.00157

Description: Comprehensive survey examining mathematical reasoning capabilities of LLMs across diverse problem types and datasets. Systematic analysis of mathematical problems, LLM-oriented techniques, factors affecting performance, and persisting challenges. arXiv

Key Characteristics:

- Most comprehensive survey of mathematical reasoning in LLMs (arXiv)
- Systematic taxonomy of problem types and solution approaches
- Analysis of factors influencing arithmetic vs. reasoning capabilities (Stanford) (arXiv)
- Identifies key challenges: calculation errors, consistency issues, brittleness (Stanford)

GSM-Plus: Comprehensive Benchmark for Evaluating Robustness

Minghao Li et al. (2024)
https://arxiv.org/abs/2402.19255

Description: Adversarial benchmark created by perturbing GSM8K problems across five perspectives: numerical variation, arithmetic variation, problem understanding, distractor insertion, and critical thinking. Evaluates whether LLMs truly understand mathematical reasoning. (arXiv)

Key Characteristics:

- Tests robustness of mathematical reasoning capabilities
- Accuracy gaps up to 20% between original GSM8K and GSM-Plus variants
- Five-dimensional perturbation framework based on Polya's principles
- Reveals gaps between benchmark performance and true understanding (arXiv)

ReaLMistake: Evaluating LLMs at Detecting Errors

Ryo Kamoi et al. (2024)
https://arxiv.org/abs/2404.03602

Description: First comprehensive error detection benchmark with objective, realistic, and diverse errors made by LLMs. Contains three challenging tasks with errors in four categories: reasoning correctness, instruction-following, context-faithfulness, and parameterized knowledge. (OpenReview)

Key Characteristics:

- First benchmark specifically for error detection in LLM responses
- Naturally occurring, not artificially generated, errors
- Objective, unambiguous error annotations by experts
- Even GPT-4 and Claude 3 Opus struggle with error detection (OpenReview)

MathEval: Comprehensive Benchmark

Zhenzhong Li et al. (2023)
📄 OpenReview: https://openreview.net/forum?id=DexGnh0EcB

Description: Comprehensive benchmark amalgamating 19 datasets spanning mathematical domains, languages (English and Chinese), problem types, and difficulty levels. Features contamination detection methodology. (OpenReview)

Key Characteristics:

- Most comprehensive multi-dataset mathematical reasoning benchmark
- Cross-lingual evaluation (English and Chinese)
- Built-in contamination detection
- Annually updated problems ensure benchmark freshness

Error Analysis and Reasoning Methods

Evaluating Mathematical Reasoning: Error Identification and Correction

Jiayu Wu et al. (2024)
📄 https://arxiv.org/abs/2406.00755

Description: Introduces EIC-Math dataset with four evaluation tasks: Error-Presence Identification, Error-Type Identification, Error-Correction, and Error-Specific Correction. Uses GPT-4 to convert ground-truth solutions into wrong solutions. (arXiv)

Key Characteristics:

- Novel evaluation perspective: examiner vs. examinee capabilities
- Four fine-grained evaluation tasks for comprehensive assessment
- Providing error type information improves correction accuracy by 45-48% (arXiv)

Can LLMs Learn from Previous Mistakes? (CoTErrorSet)

Tian Li et al. (2024)
📄 https://arxiv.org/abs/2403.20046
📄 ACL: https://aclanthology.org/2024.acl-long.169.pdf

Description: Introduces CoTErrorSet with 558,960 questions featuring both correct and incorrect reasoning chains with error type annotations. Proposes self-rethinking prompting and mistake tuning methods. (arXiv)

Key Characteristics:

- Largest-scale error-annotated reasoning dataset
- Systematic taxonomy of error types in LLM reasoning
- Enables learning from mistakes during training and inference
- Consistent performance improvements from mistake-aware methods

---

RE-IMAGINE: Symbolic Benchmark Synthesis
Xiaoxuan Xu et al. (2024)
📄 https://www.microsoft.com/en-us/research/publication/re-imagine-symbolic-benchmark-synthesis-for-reasoning-evaluation/

Description: Framework for characterizing hierarchy of reasoning abilities based on ladder of causation (associations, interventions, counterfactuals). Automated pipeline generates problem variations at different reasoning levels. Microsoft Microsoft

Key Characteristics:

- Addresses test set memorization concerns
- Three-level hierarchy: observe, mutate, counterfactual Microsoft
- Domain-agnostic framework applicable across reasoning tasks Microsoft
- Generates infinite variations not solvable by memorization

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models
Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. (2022)
📄 NeurIPS 2022

Description: Foundational paper introducing Chain-of-Thought (CoT) prompting, enabling LLMs to generate intermediate reasoning steps before producing final answers. Pathmind Mgubaidullin Demonstrates that prompting models to show reasoning significantly improves performance on complex reasoning tasks. arXiv arxiv

Key Characteristics:

- Foundational methodology for eliciting reasoning in LLMs
- Demonstrates scaling properties of reasoning capabilities
- Enables interpretation of model reasoning process
- Shows reasoning as emergent capability in large models

Arithmetic Reasoning with LLM: Prolog Generation & Permutation

Xiaocheng Yang, Brian Chen, Yik Cheung Tam (2024)
https://arxiv.org/abs/2405.17893
ACL: https://aclanthology.org/2024.naacl-short.61/

Description: Investigates using LLMs to generate Prolog programs for solving arithmetic reasoning problems. Introduces GSM8K-Prolog dataset with Prolog code solutions and predicate permutation as data augmentation. arXiv +2

Key Characteristics:

- Novel approach through logic programming arXiv
- Demonstrates separation of symbolic reasoning from calculation ACL Anthology arXiv
- Shows systematic improvement over Chain-of-Thought methods arXiv +2

---

4. RLHF and Alignment

Foundational RLHF Papers

Deep reinforcement learning from human preferences

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, Dario Amodei (2017)
NeurIPS 2017: https://arxiv.org/abs/1706.03741

Relevance: Establishes foundational framework for RLHF by demonstrating that AI systems can learn complex behaviors from non-expert human preferences over trajectory pairs. arXiv Springer Enables alignment by training reward model on human comparisons and using it to guide policy optimization.

Key Methods:

- Preference-based reward learning: trains neural network reward models to predict which trajectory segments humans prefer arXiv
- Reduces human oversight requirements by orders of magnitude compared to learning from demonstrations arXiv
- Successfully applied to Atari games and simulated robot locomotion with ~900 human comparisons (~1 hour) arXiv

---

Training language models to follow instructions with human feedback (InstructGPT)

Long Ouyang, Jeff Wu, Xu Jiang, et al. (2022)
NeurIPS 2022: https://arxiv.org/abs/2203.02155

Relevance: Demonstrates that RLHF can align large language models with user intent across diverse tasks. arXiv OpenReview Establishes the three-step alignment procedure (supervised fine-tuning, reward model training, and PPO-based RL) that became the industry standard.

Key Methods:

- Three-stage training pipeline: (1) Supervised fine-tuning on labeler demonstrations, (2) Training reward model on human preference rankings, (3) PPO reinforcement learning ResearchGate
- 1.3B parameter InstructGPT preferred to 175B GPT-3 outputs 71% of the time arXiv ResearchGate
- Improvements in truthfulness and safety while maintaining performance on public NLP benchmarks

---

Constitutional AI: Harmlessness from AI Feedback
Yuntao Bai, Saurav Kadavath, Sandipan Kundu, et al. (2022)
https://arxiv.org/abs/2212.08073

Relevance: Introduces scalable alignment method training harmless AI assistants through self-improvement guided by principles (a "constitution"). arXiv Stanford Combines supervised learning (with AI self-critique and revision) and RL from AI Feedback (RLAIF), where AI evaluates its own outputs. Stanford

Key Methods:

- Two-phase training: (1) Supervised learning with self-critiques and revisions based on constitutional principles, (2) RL phase using AI-generated preference labels (RLAIF) Stanford
- Achieves harmlessness with only ~10 human-written principles as input
- Leverages CoT-style reasoning in both critique generation and preference evaluation Stanford

---

Related Alignment Papers

Learning to summarize from human feedback
Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, Paul Christiano (2020)
NeurIPS 2020: https://arxiv.org/abs/2009.01325

Relevance: Demonstrates training language models to optimize for human preferences rather than proxy metrics like ROUGE significantly improves summary quality. NeurIPS arXiv Establishes RLHF framework for NLP tasks. arXiv

Key Methods:

- Collected large-scale dataset of human comparisons between summaries NeurIPS
- Models trained with RLHF significantly outperform human reference summaries NeurIPS

- Reward models trained on one dataset transfer to others without fine-tuning (NeurIPS)

Direct Preference Optimization: Your Language Model is Secretly a Reward Model
Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, Chelsea Finn (2023)
NeurIPS 2023: https://arxiv.org/abs/2305.18290

Relevance: Introduces simpler alternative to RLHF that eliminates need for explicit reward modeling and reinforcement learning by directly optimizing language model on preference data using classification loss.

Key Methods:

- Derives closed-form optimal policy reparameterization allowing extraction of optimal policy with only binary cross-entropy loss
- Eliminates complexity of reward model training and RL sampling during fine-tuning
- Matches/exceeds PPO-based RLHF in sentiment control, summarization, and dialogue

---

Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback
Yuntao Bai, Andy Jones, Kamal Ndousse, et al. (2022)
https://arxiv.org/abs/2204.05862

Relevance: Provides detailed documentation of applying RLHF to balance multiple objectives (helpfulness and harmlessness). Demonstrates alignment training improves performance on almost all NLP evaluations without compromising specialized skills.

Key Methods:

- Multi-objective alignment: trains models to be both helpful and harmless by mixing preference data
- Iterated online RLHF: weekly updates to preference models and RL policies
- RLHF-trained models perform better than raw models on virtually all evaluations

---

5. Evaluation Frameworks

Reasoning Trace Evaluation

Evaluating Step-by-step Reasoning Traces: A Survey
Jinu Lee, Julia Hockenmaier (2025)
https://arxiv.org/abs/2502.12289

Description: Comprehensive survey proposing taxonomy of evaluation criteria with four top-level categories: Groundedness, Validity, Coherence, and Utility. Systematically categorizes existing metrics and evaluator implementations.

Key Metrics:

- Six metric implementation types: Rule-based matching, uncertainty measures, V-information, cross-encoders, process reward models, critic models (LLM-as-a-judge), generative verifiers
- Meta-evaluation benchmarks: REVEAL, PRMBench, ProcessBench
- Demonstrates validity and coherence are highly transferable ($\varrho=0.88$)

Direct Evaluation of Chain-of-Thought in Multi-hop Reasoning with Knowledge Graphs

Minh-Vuong Nguyen, Linhao Luo, Fatemeh Shiri, et al. (2024)
📄 ACL 2024 Findings: https://arxiv.org/abs/2402.11199

Description: Proposes discriminative and generative CoT evaluation paradigms using knowledge graphs as ground truth. Measures both answer accuracy and reasoning faithfulness separately.

Key Metrics:

- Discriminative evaluation: Binary classification of reasoning step correctness using KG verification
- Generative evaluation: Entity linking and relation extraction to compare generated reasoning against KG paths
- Finds models achieve correct answers with incorrect reasoning in 30-40% of cases

---

Prompting Strategy Comparison

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. (2022)
📄 NeurIPS 2022: https://arxiv.org/abs/2201.11903

Description: Introduces chain-of-thought prompting and evaluates across arithmetic, commonsense, and symbolic reasoning tasks. Compares CoT prompting against standard prompting using few-shot exemplars.

Key Metrics:

- Answer accuracy on benchmarks: GSM8K, SVAMP, MAWPS, CommonsenseQA, StrategyQA
- Ablation comparisons: Standard prompting vs. CoT vs. equation-only vs. variable compute
- Emergence analysis: CoT reasoning emerges only at sufficient model scale (>100B parameters)

Self-Consistency Improves Chain of Thought Reasoning in Language Models
Xuezhi Wang, Jason Wei, Dale Schuurmans, et al. (2023)
ICLR 2023: https://arxiv.org/abs/2203.11171

Description: Introduces self-consistency as alternative to greedy decoding for CoT prompting. Samples diverse reasoning paths and selects most consistent answer by marginalizing over sampled paths.

Key Metrics:

- Sample-and-marginalize decoding: Generate multiple (5-40) reasoning paths and select majority answer
- Performance gains: GSM8K (+17.9%), SVAMP (+11.0%), AQuA (+12.2%), StrategyQA (+6.4%)
- Robustness analysis: Testing with varying numbers of sampled paths and temperatures

---

Human Evaluation Protocols

A Framework for Human Evaluation of Large Language Models in Healthcare (QUEST)
Thomas Yu Chow Tam et al. (2024)
https://arxiv.org/abs/2405.02559

Description: Systematic literature review of 142 studies on human evaluation of LLMs in healthcare, proposing QUEST framework with five principles. Reviews evaluation tools including Likert scales, binary correctness variables, and error categorization.

Key Protocols:

- QUEST Principles: (1) Quality of Information, (2) Understanding and Reasoning, (3) Expression Style, (4) Safety and Harm, (5) Trust and Confidence
- Evaluator selection: Guidelines for determining appropriate number of evaluators (typically 2-5 domain experts)
- Statistical measures: Inter-rater reliability (Cohen's kappa, Fleiss' kappa), correlation analysis

---

An Empirical Evaluation of Prompting Strategies for Large Language Models
Sonish Sivarajkumar, Mark Kelley, Alina Samolyk-Mazzanti, Shyam Visweswaran, Yanshan Wang (2024)
https://arxiv.org/abs/2309.08008

Description: Comprehensive experimental study comparing six prompt types (simple prefix, simple cloze, chain-of-thought, anticipatory, heuristic, ensemble) across five clinical NLP tasks.

Key Metrics:

- Accuracy as primary metric: Proportion of correct outputs

- Zero-shot vs. few-shot comparison: 2-shot examples provide consistent gains
- Heuristic and chain-of-thought prompts achieve 0.94-0.96 accuracy on clinical sense disambiguation

---

Automated Reasoning Metrics

Understanding Chain-of-Thought in LLMs through Information Theory

Jean-Francois Ton et al. (2024)
https://arxiv.org/abs/2411.11984

Description: Formalizes CoT reasoning through information-theoretic lens by quantifying "information-gain" at each reasoning step. Enables identification of failure modes without expensive annotated datasets.

Key Metrics:

- Information-gain metric: Measures how much information each step adds toward solving the problem
- False positive reduction: Significantly outperforms outcome-based methods in detecting errors
- Automated error detection: Identifies different types of errors (calculation, logic, redundancy) without human annotations

---

OCEAN: Offline Chain-of-thought Evaluation and Alignment

Junda Wu, Xintong Li, Ruoyu Wang, et al. (2024)
https://arxiv.org/abs/2410.23703

Description: Proposes offline evaluation framework modeling CoT reasoning as Markov Decision Process with knowledge graph preference modeling. Uses external knowledge graphs (Wikidata5m) to provide feedback.

Key Metrics:

- KG-based policy alignment: Models KG policy generating token-level likelihood distributions for reasoning paths
- MDP formulation: States represent reasoning steps, actions are token generations, rewards measure KG consistency
- On-policy exploration: Uses RL to model KG reasoning preference

---

Evaluating Mathematical Reasoning Beyond Accuracy (ReasonEval)

https://arxiv.org/html/2404.05692v1

Description: Introduces ReasonEval suite of evaluation metrics and LLM-based evaluators for mathematical reasoning beyond final answer accuracy. Emphasizes validity and redundancy of each reasoning step.

Key Metrics:

- Validity scoring: Step contains no mistakes in calculation and logic
- Redundancy detection: Identifying steps that lack utility but are technically correct
- Three-way classification: Categorizing each step as positive (valid + useful), negative (invalid), or redundant

---

6. GitHub Repositories

Process Supervision Implementations

OpenAI PRM800K

🔗 https://github.com/openai/prm800k

🤗 https://huggingface.co/datasets/openai/prm800k

Description: Landmark dataset for process supervision containing 800,000 step-level correctness labels on model-generated solutions to MATH problems. Includes human annotation data, grading logic for mathematical answers.

Related Paper: "Let's Verify Step by Step" (Lightman et al., 2023)

Key Features:

- 800K step-level labels with human annotation instructions
- Evaluation scripts for ORM vs PRM comparison
- Grading logic using sympy for mathematical answer verification

---

Qwen2.5-Math-PRM

🔗 https://github.com/QwenLM/Qwen2.5-Math

🤗 HuggingFace Models: Qwen2.5-Math-PRM-7B and Qwen2.5-Math-PRM-72B

Description: State-of-the-art open-source process reward models from Alibaba's Qwen team. Provide step-by-step feedback on mathematical reasoning quality and intermediate steps.

Related Paper: "The Lessons of Developing Process Reward Models in Mathematical Reasoning" (Zhang et al., 2025)

Key Features:

- Automatic step-level error annotation
- Compatible with formal verification tools

- ProcessBench benchmark (3,400 Olympiad-level problems)
- Outperforms other 7B PRMs

RLHFlow/RLHF-Reward-Modeling

🔗 https://github.com/RLHFlow/RLHF-Reward-Modeling

Description: Comprehensive recipes for training reward models including Bradley-Terry models, pairwise preference models, multi-objective reward models (ArmoRM), and process-supervised reward models (PRM/ORM). Achieved #1 ranking on RewardBench.

Related Papers: "RLHF Workflow: From Reward Modeling to Online RLHF" (Dong et al., 2024)

Key Features:

- Multiple RM architectures
- State-of-the-art performance (ArmoRM-Llama3-8B ranked #1 on RewardBench)
- Training code for PRMs and ORMs
- Supports 4xA40 or 4xA100 training

Reasoning Evaluation Frameworks

LLM Reasoners (Maitrix-org)

🔗 https://github.com/maitrix-org/llm-reasoners

Description: Comprehensive library for advanced LLM reasoning algorithms including MCTS, Tree-of-Thoughts, Chain-of-Thought, RAP, PRM-guided search. Features intuitive visualization tools and SGLang integration for 100x speedup.

Related Paper: "LLM Reasoners: New Evaluation, Library, and Analysis of Step-by-Step Reasoning" (Hao et al., 2024, COLM 2024)

Key Features:

- 15+ reasoning algorithms
- Visualization tool for MCTS/reasoning processes
- SGLang integration for inference-time scaling with PRMs
- ReasonerAgent for web research

Chain-of-Thought Hub

🔗 https://github.com/FranxYao/chain-of-thought-hub

Description: Benchmarking platform for evaluating LLMs' complex reasoning abilities with chain-of-thought prompting across multiple challenging tasks. Includes GSM8K, MATH, MMLU, BBH, HumanEval, C-Eval.

Key Features:

- Covers 9+ benchmarks
- Evaluation scripts for GPT/Claude/LLaMA/Falcon
- Standardized prompts library
- Tracks performance of 30+ models

Awesome-LLM-Reasoning

🔗 https://github.com/atfortes/Awesome-LLM-Reasoning

Description: Comprehensive collection of papers and resources on LLM reasoning from Chain-of-Thought to OpenAI o1 and DeepSeek-R1. Covers reasoning algorithms, benchmarks, datasets, training methods.

Key Features:

- 200+ papers on reasoning
- Covers inference-time compute scaling
- Process supervision papers
- Systematic categorization by method type

Math Reasoning Datasets with Code

OpenAI GSM8K

🔗 https://github.com/openai/grade-school-math
🤗 https://huggingface.co/datasets/openai/gsm8k

Description: Grade School Math 8K dataset with 8.5K high-quality grade school math word problems. Problems require 2-8 steps using basic arithmetic.

Related Paper: "Training Verifiers to Solve Math Word Problems" (Cobbe et al., 2021)

Key Features:

- 8.5K human-written problems
- Natural language solutions with calculator annotations

- Evaluation code and example model solutions
- MIT licensed

MATH Dataset (Hendrycks)
🔗 https://github.com/hendrycks/math
🤗 https://huggingface.co/datasets/hendrycks/competition_math

Description: Competition-level mathematics dataset with 12K problems from AMC 10, AMC 12, AIME competitions across 7 categories. Significantly more challenging than GSM8K.

Related Paper: "Measuring Mathematical Problem Solving With the MATH Dataset" (Hendrycks et al., 2021, NeurIPS)

Key Features:

- 12K competition-level problems
- Step-by-step solutions in LaTeX
- 5 difficulty levels
- Dataset loaders and evaluation code

Google DeepMind Mathematics Dataset
🔗 https://github.com/google-deepmind/mathematics_dataset

Description: Synthetic mathematics dataset generator producing question-answer pairs across school-level difficulty. Covers algebra, arithmetic, calculus, comparison, measurement, numbers, polynomials, probability.

Key Features:

- Programmatic generation
- Curriculum learning support (easy/medium/hard splits)
- Diverse question types
- 40+ mathematical categories

Qwen2.5-Math Complete Suite
🔗 https://github.com/QwenLM/Qwen2.5-Math

Description: Comprehensive math-specific LLM series including base models (1.5B/7B/72B), instruction-tuned models, and reward models. Supports both Chain-of-Thought and Tool-integrated Reasoning (TIR).

Key Features:

- Multiple model sizes
- Bilingual support (Chinese and English)
- TIR with Python interpreter
- Qwen2.5-Math-72B-Instruct achieves 92.9% on MATH with RM@8

LLM Reasoning Benchmarking Tools

EleutherAI LM Evaluation Harness

🔗 https://github.com/EleutherAI/lm-evaluation-harness

Description: Framework for few-shot evaluation of language models across hundreds of tasks. Includes GSM8K, MATH, BBH, MMLU and many other reasoning benchmarks.

Key Features:

- 200+ task implementations
- Standardized metrics
- CoT evaluation support
- Various prompting strategies

LiveBench

🔗 https://github.com/LiveBench/LiveBench

Description: Contamination-free LLM benchmark that releases new questions monthly based on recent datasets, arXiv papers, news, and IMDb synopses. Contains 18 diverse tasks across 6 categories.

Key Features:

- Monthly updates to prevent contamination
- Objective ground-truth answers
- 18 reasoning tasks
- Automatic evaluation without LLM judges

Awesome System-2 Reasoning

🔗 https://github.com/zzli2022/Awesome-System2-Reasoning-LLM

Key Features:

- Curated paper collection on reasoning models
- Tracks o1/R1 developments
- Inference-time scaling methods
- Process verification techniques

Reward Model Implementations

Self-Rewarding LM (lucidrains)

🔗 https://github.com/lucidrains/self-rewarding-lm-pytorch

Description: Implementation of Self-Rewarding Language Models from MetaAI where models act as their own reward model. Supports iterative training with DPO.

Related Paper: "Self-Rewarding Language Models" (Meta AI, 2024)

Key Features:

- Self-reward via LLM-as-a-Judge
- Flexible training configurations
- Supports SPIN and DPO
- Iterative self-improvement

Awesome Reward Models

🔗 https://github.com/JLZhong23/awesome-reward-models

Description: Comprehensive collection of papers and resources on reward models including outcome-based (ORM) and process-based (PRM) reward models.

Key Features:

- 100+ papers on reward modeling
- Categorized by type (ORM/PRM)
- Inference-time scaling papers
- LLM-as-judge methods

PURE (Process Reward Model Research)

🔗 https://github.com/CJReinforce/PURE

Description: Official code for "Stop Summation: Min-Form Credit Assignment Is All Process Reward Model Needs for Reasoning." Trains PRM using PRM800K dataset and Qwen2.5-Math-7B, achieving 82.6% on MATH500.

Key Features:

- Efficient PRM training
- 1/50th RL data requirement
- Implements PURE-VR and PURE-PRM+VR
- Achieves SOTA with limited resources

Step-DPO

🔗 https://github.com/dvlab-research/Step-DPO

Description: Step-wise Preference Optimization for long-chain reasoning. Constructs 10K step-wise preference pairs and applies DPO at the step level.

Related Paper: "Step-DPO: Step-wise Preference Optimization for Long-chain Reasoning" (Lai et al., 2024)

Key Features:

- 10K step-wise preference dataset
- Data construction pipeline
- Significant improvements with minimal data
- Works on Qwen2/Llama-3/DeepSeekMath

Chain-of-Thought Prompting Libraries

Chain-of-Thought Papers (Timothyxxx)

🔗 https://github.com/Timothyxxx/Chain-of-ThoughtsPapers

Description: Comprehensive collection of papers starting from "Chain of Thought Prompting Elicits Reasoning in Large Language Models." Includes 100+ papers on CoT variants.

Key Features:

- Chronological organization

- Links to implementations
- Covers CoT variants and extensions
- Tool-use and environment interaction

Auto-CoT (Amazon Science)
🔗 https://github.com/amazon-science/auto-cot

Description: Official implementation of Automatic Chain of Thought Prompting. Automatically generates diverse CoT demonstrations without manual prompt engineering.

Related Paper: "Automatic Chain of Thought Prompting in Large Language Models" (Zhang et al., ICLR 2023)

Key Features:

- Automatic demonstration selection
- Diversity-based sampling
- Eliminates manual prompt engineering
- Reproduction scripts for multiple benchmarks

Active Prompting (shizhediao)
🔗 https://github.com/shizhediao/active-prompt

Description: Active learning approach for selecting the most helpful questions to annotate with CoT demonstrations. Uses uncertainty metrics to identify valuable examples.

Key Features:

- Uncertainty-based question selection
- Active learning for prompting
- 7% improvement on GSM8K
- Multiple uncertainty metrics

TRL (Transformers Reinforcement Learning)
🔗 https://huggingface.co/docs/trl/en/prm_trainer

Description: HuggingFace library that includes PRMTrainer for training process reward models. Provides easy interface for taking an LLM and PRM-format dataset.

Key Features:

- PRMTrainer class
- Integrates with HuggingFace
- Supports Math-Shepherd format
- Token-level supervision

Annotation Tools

OpenAI PRM800K Annotation Interface
🔗 https://github.com/openai/prm800k (includes annotation instructions)

Description: The annotation instructions and interface design used by OpenAI for collecting 800K step-level labels. Includes Phase 1 and Phase 2 labeling guidelines.

Key Features:

- Two-phase annotation protocol
- Quality control with gold labels
- Labeler screening system
- Documents human annotation process

---

7. Legal/Regulatory AI

Legal Frameworks and AI Explainability

Accountability of AI Under the Law: The Role of Explanation
Finale Doshi-Velez, Mason Kortz, Ryan Budish, et al. (2017)
📄 Berkman Klein Center: https://arxiv.org/pdf/1711.01134

Relevance: Interdisciplinary collaboration between legal and computer science scholars examining when and what kind of explanation might be required of AI systems, with explicit focus on "right to explanation" debated in EU's GDPR.

Key Regulatory Implications:

- Establishes legally-operative explanations from AI are technically feasible using local explanation and counterfactual faithfulness
- Documents that administrative agencies in U.S. are legally required to explain algorithmic decisions under Administrative Procedure Act

- Demonstrates GDPR Articles 13-15 and 22 create substantive rights to "meaningful information about the logic involved"

Explainable AI and Law: An Evidential Survey
K.M. Richmond, S.M. Muddamsetty, T. Gammeltoft-Hansen, H.P. Olsen, T.B. Moeslund (2023)
📄 Digital Society: https://link.springer.com/article/10.1007/s44206-023-00081-z

Relevance: Systematic survey establishing novel taxonomy linking different forms of legal inference in specific legal sub-domains to specific forms of algorithmic decision-making. Surveys multiple national jurisdictions.

Key Regulatory Implications:

- Provides evidence that current legal frameworks across surveyed jurisdictions don't extend into causal explainability requirements
- Establishes "black-box" nature of AI creates particular accountability obstacles in legal contexts
- Demonstrates heterogeneity in legal logics requires tailored XAI approaches

---

The Judicial Demand for Explainable Artificial Intelligence
Yavar Bathaee (2019)
📄 Columbia Law Review: https://www.columbialawreview.org/content/the-judicial-demand-for-explainable-artificial-intelligence/

Relevance: Argues that judges confronting machine learning algorithms with increasing frequency should demand explanations for algorithmic outcomes. Positions judges as key actors who will shape the nature and form of XAI through common law development.

Key Regulatory Implications:

- Establishes judges have both incentive and authority to demand explanations
- Common law development by courts provides advantages over statutory approaches
- Courts will create "common law of xAI" sensitive to different audiences and contexts

---

The Right to Explanation, Explained
Margot E. Kaminski (2019)
📄 Berkeley Technology Law Journal: https://btlj.org/data/articles2019/34_1/05_Kaminski_Web.pdf

Relevance: Provides definitive clarification of GDPR's algorithmic accountability regime, incorporating authoritative Article 29 Working Party/European Data Protection Board guidelines. Demonstrates GDPR creates broader, stronger, and deeper algorithmic accountability system.

Key Regulatory Implications:

- Article 22 is prohibition on automated decision-making (not merely right to object)
- GDPR creates "qualified transparency" regime with different depths of disclosure
- Working Party guidelines narrow exceptions, close trade secrets loopholes

---

Emerging Regulatory Frameworks

Argumentation-Based Explainability for Legal AI: Comparative and Regulatory Perspectives

Andrada Iulia Prajescu, Roberto Confalonieri (2024)
https://arxiv.org/abs/2510.11079

Relevance: Analyzes how various XAI methods align with emerging regulatory frameworks, particularly EU GDPR and AI Act. Argues computational argumentation frameworks best positioned to provide legally relevant explanations.

Key Regulatory Implications:

- Argumentation-based approaches to XAI better align with legal reasoning requirements
- Identifies open challenges including bias mitigation, empirical validation
- Meeting both technical and normative requirements requires approaches handling defeasibility

---

AI Auditing: First Steps Towards the Effective Regulation of Artificial Intelligence Systems

Edwin A. Farley, Christian R. Lansang (2025)
Harvard Journal of Law & Technology: https://jolt.law.harvard.edu/assets/digestImages/Farley-Lansang-AI-Auditing-publication-2.13.2025.pdf

Relevance: Proposes government-mandated AI audits conducted by professional auditors following established standards. Draws insights from financial auditing and proposes three-component audit structure (data, model, deployment).

Key Regulatory Implications:

- AI auditing should target three components: data collection/management, model development, deployment
- Oversight agencies can use "technology forcing" policies to advance auditing technology
- Auditing creates "virtuous cycle" bolstering auditor independence and public understanding

---

The Role of Explainable AI in the Context of the AI Act

Forough Poursabzi-Sangdeh, Daniel G. Goldstein (2023)

📄 FAccT 2023: https://dl.acm.org/doi/10.1145/3593013.3594069

Relevance: Provides interdisciplinary analysis of how EU AI Act addresses opaque AI systems, examining whether Act mandates explainable AI for high-risk systems. Clarifies debates about whether black-box AI models must be replaced.

Key Regulatory Implications:

- AI Act doesn't mandate XAI as technical requirement, nor ban black-box systems
- Requires "sufficient transparency to enable users to interpret the system's output"
- Identifies tensions between need for meaningful explanations and intellectual property risks

---

Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise
Bryan Casey, Ashkon Farhangi, Roland Vogl (2019)
📄 Berkeley Technology Law Journal: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3143325

Relevance: Argues GDPR debate over "right to explanation" has overshadowed most revolutionary change: sweeping new enforcement powers given to European Data Protection Authorities. Positions algorithmic audits as key compliance mechanism.

Key Regulatory Implications:

- GDPR Chapters 6 and 8 grant DPAs potent investigatory, advisory, corrective, and punitive powers
- GDPR's true power derives from synergistic effects combining right to explanation with algorithmic auditing
- Algorithmic audits provide more muscular accountability than individual explanations alone

---

Summary and Key Trends

Process Supervision Evolution

2021-2023: Foundation and Validation

- Early work established outcome reward models (Cobbe et al. 2021)
- First comprehensive comparison showed process supervision improves reasoning quality (Uesato et al. 2022)
- Large-scale human feedback demonstrated clear superiority of process supervision (Lightman et al. 2023, PRM800K)

2024-2025: Automation and Scalability

- Major focus shifted to automating process supervision data collection (OmegaPRM, Math-Shepherd)
- Methods using MCTS, Monte Carlo estimation, and LLM-as-a-judge emerged
- OmegaPRM collected 1.5M+ process labels automatically

2025: Data Efficiency and Generative Approaches

- ThinkPRM demonstrated generative PRMs outperform discriminative models with 100× less training data
- Process-based self-rewarding works for reasoning when combined with step-wise supervision (Zhang et al. 2025)
- Shift toward verbalized, interpretable process rewards

Interpretability and Faithfulness Challenges

- **Inverse scaling problem**: Larger models often less faithful (Lanham et al. 2023)
- **Distinction between self-consistency and true faithfulness** (Parcalabescu & Frank 2024)
- **Causal analysis** emerging as rigorous measurement approach (Paul et al. 2024)
- **Production models still exhibit unfaithfulness**: GPT-4o-mini (13%), Haiku 3.5 (7%) (Arcuschin et al. 2025)
- **Formal verification** complements neural approaches (Zhou & He 2025)

Legal and Regulatory Landscape

- **Technical feasibility exists**: AI systems can provide legally-operative explanations without revealing trade secrets
- **Regulatory frameworks converging**: GDPR, AI Act, and national laws increasingly require explainability/transparency
- **Multiple accountability mechanisms**: Beyond individual explanations—algorithmic auditing, impact assessments, human oversight
- **Judicial role emerging**: Courts developing common law of XAI through case-by-case adjudication
- **Process supervision connection**: Human oversight, auditing, and continuous monitoring requirements embed accountability throughout AI system lifecycle

Practical Applications

Verification/Best-of-N Selection: PRMs enable more effective reranking of multiple candidate solutions at inference time

Reinforcement Learning: Process supervision provides denser, more informative training signals for RL-based finetuning

Interpretability: Process-level feedback provides transparency into reasoning steps, making it easier to identify and correct errors

Alignment: Process supervision directly rewards models for following human-endorsed reasoning chains

Educational Applications: Step-by-step verification enables better pedagogical applications where reasoning process matters

Legal Compliance: Process supervision and interpretability connect to regulatory requirements for explainable and auditable AI systems

---

Priority Resources for Your Project

Must-Read Papers

1. Uesato et al. 2022 - First process vs. outcome comparison
2. Lightman et al. 2023 - PRM800K and state-of-the-art process supervision
3. Zhang et al. 2025 - Process-based self-rewarding
4. Lanham et al. 2023 - Measuring faithfulness in CoT
5. Christiano et al. 2017 - RLHF foundations
6. Ouyang et al. 2022 - InstructGPT
7. Doshi-Velez et al. 2017 - Accountability of AI under the law

Essential Datasets

1. PRM800K - 800K step-level labels
2. GSM8K - 8.5K grade school math problems
3. MATH - 12.5K competition-level problems
4. ProcessBench - 3.4K Olympiad-level problems

Key GitHub Repositories

1. openai/prm800k - Process supervision dataset
2. QwenLM/Qwen2.5-Math - State-of-the-art PRMs
3. RLHFlow/RLHF-Reward-Modeling - Comprehensive RM training
4. maitrix-org/llm-reasoners - Reasoning algorithms library

5. FranxYao/chain-of-thought-hub - Benchmarking platform

This research compilation covers 70+ papers and 25+ GitHub repositories providing comprehensive foundation for work on process vs. outcome supervision for trustworthy and interpretable LLM reasoning, with connections to technical research, evaluation methodologies, and legal/regulatory requirements.