Novelty Assessment Report

Paper: Foundational Automatic Evaluators: Scaling Multi-Task Generative Evaluator Training for Reasoning-Centric Domains
PDF URL: https://openreview.net/pdf?id=89Ei7PVpNl
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Finetuning specialized generative evaluators has emerged as a popular paradigm to meet the increasing demand for scalable evaluation during both training and test-time. However, recent work has largely focused on applying new methodology, such as reinforcement learning (RL), to training evaluators, shying away from large-scale, data-driven development. In this work, we focus on data scaling, curating a set of 2.5M samples spanning five unique evaluation tasks (pairwise, step-level, reference-free and reference-based verification, and single rating) and multiple domains focused on reasoning evaluation. With our data, we train Foundational Automatic Reasoning Evaluators (FARE), a family of 8B and 20B (with 3.6B active) parameter evaluators, with a simple iterative rejection-sampling supervised finetuning (SFT) approach. FARE-8B challenges larger specialized RL-trained evaluators and FARE-20B sets the new standard for open-source evaluators, surpassing specialized 70B+ evaluators. Beyond static benchmarks, we evaluate FARE in real-world tasks: As inference-time rerankers, FARE-20B achieves near-oracle performance on MATH. As verifiers in RL training, FARE improves the downstream RL-trained model performance by up to 14.1\% vs. string-matching verifiers. When initialized from FARE, a continually-finetuned FARE-Code outperforms gpt-oss-20B by 65% on evaluating test-case quality

Core Task Landscape

This paper addresses: **training multi-task generative evaluators for reasoning evaluation**

A total of **43 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Generative Evaluator Architecture and Training**
- **Reasoning Evaluation Methodologies and Benchmarks**
- **Reasoning Systems and Generative Models for Complex Tasks**
- **Multi-Task Learning and Generative Model Applications**
- **Domain-Specific Generative AI Applications**

Complete Taxonomy Tree

- training multi-task generative evaluators for reasoning evaluation Survey Taxonomy
- Generative Evaluator Architecture and Training
  - Multi-Task Evaluator Training Frameworks ★ (3 papers)
  - [0] Foundational Automatic Evaluators: Scaling Multi-Task Generative Evaluator Training for Reasoning-Centric Domains (Anon et al., 2026) View paper
  - [1] Praetor: A Fine-Grained Generative LLM Evaluator with Instance-Level Customizable Evaluation Criteria (Yongqi Leng, 2025) View paper
  - [9] J4R: Learning to Judge with Equivalent Initial State Group Relative Policy Optimization (Xu, 2025) View paper
  - Multimodal Evaluator Design (3 papers)
  - [18] Flex-Judge: Text-Only Reasoning Unleashes Zero-Shot Multimodal Evaluators (Ko, 2025) View paper
  - [19] ReFeR: Improving Evaluation and Reasoning through Hierarchy of Models (Chandra, 2024) View paper
  - [22] ARM-Thinker: Reinforcing Multimodal Generative Reward Models with Agentic Tool Use and Visual Reasoning (Shengyuan Ding, 2025) View paper
  - Test-Time Compute Scaling for Evaluation (1 papers)
  - [28] Scaling Evaluation-time Compute with Reasoning Models as Process Evaluators (Kim, 2025) View paper
- Reasoning Evaluation Methodologies and Benchmarks
  - Chain-of-Thought and Step-Level Reasoning Assessment (4 papers)
  - [2] Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models (Hao, 2024) View paper
  - [6] Direct evaluation of chain-of-thought in multi-hop reasoning with knowledge graphs (Nguyen Minh Vuong, 2024) View paper
  - [35] Chain-of-Thought Reasoning Evaluation Framework for Question Answering System (Shivani G Aithal, 2025) View paper
  - [36] SocREval: Large Language Models with the Socratic Method for Reference-Free Reasoning Evaluation (He, 2023) View paper
  - Benchmark-Free and Generative Evaluation Paradigms (2 papers)
  - [11] BeyondBench: Benchmark-Free Evaluation of Reasoning in Language Models (Srivastava Gaurav, 2025) View paper
  - [16] Generative Evaluation of Complex Reasoning in Large Language Models (Haowei Lin, 2025) View paper
  - Domain-Specific Reasoning Evaluation (4 papers)
  - [26] MMGR: Multi-Modal Generative Reasoning (Zefan Cai, 2025) View paper
  - [31] Automating Expert-Level Medical Reasoning Evaluation of Large Language Models (Zhou, 2025) View paper
  - [32] Prompting Contrastive Explanations for Commonsense Reasoning Tasks (Paranjape, 2021) View paper

    - [42] CS-NLP team at SemEval-2020 Task 4: Evaluation of State-of-the-art NLP Deep Learning Architectures on Commonsense Reasoning Task (Saeedi, 2020) View paper
    - Autonomous and Meta-Evaluation Frameworks (2 papers)
    - [29] Validating Computational Deliberation: An Empirical Analysis of Governed vs. Generative AI Reasoning (Altarkait, 2025) View paper
    - [33] Autonomous Evaluation of LLMs for Truth Maintenance and Reasoning Tasks (Karia, 2024) View paper
- Reasoning Systems and Generative Models for Complex Tasks
    - Multi-Agent and Collaborative Reasoning Systems (3 papers)
    - [10] Wireless multi-agent generative AI: From connected intelligence to collective intelligence (Zou, 2023) View paper
    - [14] MechAgents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge (Ni Bo, 2023) View paper
    - [15] Multi-Agent Generative AI: Coordinated Synthesis for Complex Problem-Solving (Chandra Sekhar Oleti, 2024) View paper
    - Neurosymbolic and Hybrid Reasoning Approaches (3 papers)
    - [23] Enhancing Large Language Models with Neurosymbolic Reasoning for Multilingual Tasks (Agrawal Ameeta, 2025) View paper
    - [25] A Generative AI-Enhanced Case-Based Reasoning Method for Risk Assessment: Ontology Modeling and Similarity Calculation Framework (Jiayi Sun, 2025) View paper
    - [39] Learning to Solve Abstract Reasoning Problems with Neurosymbolic Program Synthesis and Task Generation (Jakub Bednarek, 2024) View paper
    - Single-Agent Reasoning and Planning Models (2 papers)
    - [24] ReasonIR: Training Retrievers for Reasoning Tasks (Shao, 2025) View paper
    - [30] A Modular Multitask Reasoning Framework Integrating Spatio-temporal Models and LLMs (Ji Jiahao, 2025) View paper
    - Abstract and Visual Reasoning Models (2 papers)
    - [8] Beyond Task-Specific Reasoning: A Unified Conditional Generative Framework for Abstract Visual Reasoning (Shi Fan, 2025) View paper
    - [41] Abstract Reasoning via Logic-guided Generation (Yu, 2021) View paper
- Multi-Task Learning and Generative Model Applications
    - Multi-Task Reinforcement Learning with Generative Models (2 papers)
    - [5] Diffusion Model is an Effective Planner and Data Synthesizer for Multi-Task Reinforcement Learning (He, 2023) View paper
    - [40] Exploration for Multi-task Reinforcement Learning with Deep Generative Models (Bangaru, 2022) View paper
    - Multi-Task Supervised Learning and Generative Frameworks (3 papers)
    - [12] UniEvent: Unified Generative Model with Multi-Dimensional Prefix for Zero-Shot Event-Relational Reasoning (ZhengWei-tao, 2023) View paper
    - [37] Multi-Task Training with In-Domain Language Models for Diagnostic Reasoning (Sharma, 2023) View paper
    - [38] Multi-Task Learning of Japanese How-to Tip Machine Reading Comprehension by a Generative Model (Xiaotian Wang, 2023) View paper
    - Multi-Task Video and Vision-Language Generation (2 papers)
    - [27] ProfVLM: A Lightweight Video-Language Model for Multi-View Proficiency Estimation (Bianchi Edoardo, 2025) View paper
    - [34] FullDiT: Multi-Task Video Generative Foundation Model with Full Attention (Ju, 2025) View paper
- Domain-Specific Generative AI Applications
    - Generative AI for Education and Responsible Development (1 papers)
    - [3] Towards responsible development of generative AI for education: An evaluation-driven approach (Kunesch, 2024) View paper
    - Generative AI for Telecommunications and Network Management (1 papers)
    - [4] TAIA: Telco Generative AI-powered Multi-Agent Assistant for managing Cloud-Native Networks (Grzegorz Panek, 2025) View paper
    - Generative AI for Robotics and Autonomous Systems (2 papers)
    - [17] MTD-GPT: A Multi-Task Decision-Making GPT Model for Autonomous Driving at Unsignalized Intersections (Jiaqi Liu, 2023) View paper
    - [20] CognitiveOS: Large Multimodal Model Based System to Endow Any Type of Robot with Generative AI (Artem Lykov, 2024) View paper
    - Generative AI for Recommendation and Personalization (2 papers)
    - [7] Towards Unified Multi-Modal Personalization: Large Vision-Language Models for Generative Recommendation and Beyond (Wei, 2024) View paper
    - [13] Enhancing explainable recommendations: Integrating reason generation and rating prediction through multi-task learning (Xingyu Zhu, 2024) View paper
    - Generative AI for Specialized Professional Domains (2 papers)
    - [21] SAI4EO: Generative AI for Earth Observation Tasks (N Taggio, 2025) View paper
    - [43] Intelligent Generative Design: A New Mechanical Design Concept (Fangwei Ning, n.d.) View paper

Narrative

Core task: training multi-task generative evaluators for reasoning evaluation. This field addresses the challenge of automatically assessing complex reasoning outputs across diverse problem types, moving beyond traditional metrics toward learned evaluators that can handle multiple tasks simultaneously. The taxonomy reveals five major branches: Generative Evaluator Architecture and Training focuses on building and optimizing the evaluator models themselves, including multi-task training frameworks like those explored in Foundational Automatic Evaluators[0] and Praetor[1]; Reasoning Evaluation Methodologies and Benchmarks develops systematic approaches for measuring reasoning quality, as seen in Chain-of-Thought Evaluation[6] and Medical Reasoning Evaluation[31]; Reasoning Systems and Generative Models for Complex Tasks examines the reasoning capabilities being evaluated, from abstract reasoning in Unified Abstract Reasoning[8] to planning in Diffusion Planner[5]; Multi-Task Learning and Generative Model Applications explores broader multi-task architectures like Multi-Task Deep Generative[40] and Modular Multitask Reasoning[30]; and Domain-Specific Generative AI Applications targets specialized evaluation needs across fields from education in Responsible Generative AI Education[3] to earth observation in SAI4EO[21].

A particularly active tension exists between general-purpose multi-task evaluators and domain-specialized assessment frameworks, with works like Flex-Judge[18] and Autonomous LLM Evaluation[33] pushing toward flexible evaluation across tasks while others emphasize depth in specific reasoning types. Foundational Automatic Evaluators[0] sits squarely within the Multi-Task Evaluator Training Frameworks cluster, sharing conceptual ground with Praetor[1] and J4R[9] in developing unified training approaches for cross-task evaluation. Where Praetor[1] might emphasize particular architectural choices for multi-domain assessment, Foundational Automatic Evaluators[0] appears to establish core principles for training evaluators that generalize across reasoning types. This positioning reflects

a broader shift in the field toward treating evaluation itself as a learnable multi-task problem, rather than relying on task-specific heuristics or human annotation at scale.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Praetor: A Fine-Grained Generative LLM Evaluator with Instance-Level Customizable Evaluation Criteria

Authors: Yongqi Leng, Renren Jin, Yue Chen, Zhuowen Han, Ling Shi, et al. (10 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â , we find that improvements are lower on Coding, Math, and Reasoning tasks than other tasks. We checked the output of Praetor and observed that, for some complex questions, it was â

Relationship Analysis

Both papers belong to the Multi-Task Evaluator Training Frameworks category, focusing on training generative evaluators across multiple evaluation tasks using supervised learning approaches. They overlap in training evaluators for multiple tasks (pairwise comparison, verification, rating) across diverse domains using large-scale curated datasets with supervised finetuning. The key difference is that the original paper (FARE) emphasizes iterative rejection-sampling SFT at 2.5M samples with a focus on reasoning-centric domains and downstream RL applications, while Praetor focuses on instance-level customizable evaluation criteria with 947K samples, hierarchical guidelines, and multi-language support (Chinese and English).

2. J4R: Learning to Judge with Equivalent Initial State Group Relative Policy Optimization

Authors: Xu, Austin, Zhou, Yilun, Austin Xu, et al. (15 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

To keep pace with the increasing pace of large language models (LLM) development, model output evaluation has transitioned away from time-consuming human evaluation to automatic evaluation, where LLMs themselves are tasked with assessing and critiquing other model outputs. LLM-as-judge models are a class of generative evaluators that excel in evaluating relatively simple domains, like chat quality, but struggle in reasoning intensive domains where model responses contain more substantive and cha...

Relationship Analysis

Both papers belong to the Multi-Task Evaluator Training Frameworks category, focusing on training generative evaluators for reasoning evaluation tasks. They overlap in addressing multi-task evaluation capabilities (pairwise, step-level, verification) for reasoning-intensive domains. However, the original paper (FARE) emphasizes large-scale data curation (2.5M samples) with iterative rejection-sampling SFT across diverse reasoning domains, while the candidate paper (J4R) focuses on a novel RL algorithm (EIS-GRPO) to address positional biases in judge training, using a smaller 7B model with reinforcement learning rather than supervised approaches.

Contributions Analysis

Overall novelty summary. The paper introduces FARE, a family of generative evaluators trained on 2.5M samples spanning five evaluation tasks and multiple reasoning domains. It resides in the Multi-Task Evaluator Training Frameworks leaf, which contains only three papers including this work. This is a relatively sparse research direction within the broader taxonomy of 43 papers across 19 leaf nodes, suggesting the specific focus on large-scale data-driven multi-task evaluator training remains underexplored compared to adjacent areas like reasoning benchmarks or domain-specific applications.

The taxonomy reveals neighboring work in Multimodal Evaluator Design and Test-Time Compute Scaling for Evaluation within the same parent branch, alongside extensive activity in Reasoning Evaluation Methodologies covering step-level assessment and benchmark-free paradigms. The paper's emphasis on data scaling and supervised finetuning distinguishes it from these adjacent directions, which prioritize architectural diversity or inference-time computation. The scope_note for its leaf explicitly excludes single-task evaluators, positioning FARE as part of a push toward unified evaluation frameworks rather than specialized judges.

Among 20 candidates examined across three contributions, no clearly refuting prior work was identified. The multi-task dataset contribution examined 10 candidates with none providing overlapping prior work; the iterative rejection sampling approach examined 4 candidates with similar results; and the FARE model family examined 6 candidates without refutation. This limited search scope—20 papers from semantic search and citation expansion—suggests the analysis captures immediate neighbors but cannot claim exhaustive coverage of all potentially relevant multi-task evaluator training literature.

Given the sparse taxonomy leaf and absence of refuting candidates within the examined scope, the work appears to occupy relatively open ground in large-scale data-driven multi-task evaluator training. However, the limited search scale and the presence of only two sibling papers mean this assessment reflects local novelty within top-20 semantic matches rather than a comprehensive field survey. The taxonomy structure indicates active development in related evaluation methodologies, suggesting the broader evaluation landscape is evolving rapidly.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Multi-task, multi-domain dataset for reasoning evaluation

Description: The authors curate a dataset of 2.5 million samples spanning five evaluation tasks (pairwise, step-level, reference-free and reference-based verification, and single rating) across multiple reasoning-centric domains. This dataset combines existing high-quality annotations with newly generated synthetic data.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. TerraGen: A Unified Multi-Task Layout Generation Framework for Remote Sensing Data Augmentation

URL: View paper

Brief Assessment

TerraGen[54] focuses on remote sensing image generation for vision tasks (detection, segmentation), not reasoning evaluation. The datasets serve fundamentally different purposes—visual data augmentation versus reasoning assessment.

2. Towards Hierarchical Multi-Step Reward Models for Enhanced Reasoning in Large Language Models

URL: View paper

Brief Assessment

Hierarchical Reward Models[59] focuses on hierarchical evaluation of reasoning steps in MCTS-generated trajectories, not on curating multi-task, multi-domain datasets spanning diverse evaluation protocols (pairwise, step-level, verification, rating) across multiple domains as in the original paper.

3. SMIR: Efficient Synthetic Data Pipeline To Improve Multi-Image Reasoning

URL: View paper

Brief Assessment

SMIR[63] focuses on multi-image reasoning tasks with synthetic data generation for vision-language models, not on multi-task evaluation across diverse reasoning domains (math, code, tool-use, chat, etc.) as in the original paper.

---

4. Enhancing logical reasoning in large language models through graph-based synthetic data

URL: View paper

Brief Assessment

Graph-Based Synthetic Data[55] focuses on generating synthetic reasoning data from graph structures for specific logical reasoning tasks (kinship and spatial reasoning), not on creating a multi-task, multi-domain evaluation dataset spanning pairwise comparison, step-level evaluation, and verification tasks across diverse domains like math, code, tool-use, and chat.

---

5. Enhancing Domain-Specific Retrieval-Augmented Generation: Synthetic Data Generation and Evaluation using Reasoning Models

URL: View paper

Brief Assessment

Domain-Specific RAG Enhancement[60] focuses on synthetic QA pair generation for RAG chunking optimization in technical domains (finance, biomedical, cybersecurity), not on creating multi-task evaluation datasets for training reasoning evaluators across pairwise/step-level/verification tasks.

---

6. Promptonomyvit: Multi-task prompt learning improves video transformers using synthetic scene data

URL: View paper

Brief Assessment

Promptonomyvit[57] focuses on multi-task prompt learning for video transformers using synthetic scene data (depth, segmentation, 3D pose) to improve action recognition, not on creating datasets for reasoning evaluation with synthetic data augmentation.

---

7. Versaprm: Multi-domain process reward model via synthetic reasoning data

URL: View paper

Brief Assessment

Versaprm[58] focuses on process reward model training data for step-level reasoning evaluation across domains, not on multi-task evaluator training with five distinct evaluation tasks (pairwise, step-level, reference-free/based verification, single rating) as in the original paper.

---

8. Beyond Intelligence: The Synthetic Cognitive Augmentation Network Using Experts

URL: View paper

Brief Assessment

Synthetic Cognitive Augmentation[62] focuses on general open-domain task generation and control, not specifically on multi-task reasoning evaluation datasets. The provided context fragments do not demonstrate prior work on curating evaluation datasets spanning pairwise, step-level, and verification tasks for reasoning-centric domains.

---

9. Internbootcamp technical report: Boosting llm reasoning with verifiable task scaling

URL: View paper

Brief Assessment

Internbootcamp[61] focuses on task environments for RL training with automated generation and verification, not on curating a static evaluation dataset. The original paper curates 2.5M samples for training evaluators across five evaluation tasks, while the candidate creates interactive task environments for training reasoning models.

---

10. SHIFT: A Synthetic Driving Dataset for Continuous Multi-Task Domain Adaptation

URL: View paper

Brief Assessment

SHIFT[56] focuses on autonomous driving perception tasks with synthetic data for domain adaptation under environmental shifts (weather, time, density). The original paper addresses reasoning evaluation across math, code, and natural language domains with different evaluation task types (pairwise, verification, rating). These are fundamentally different application domains and dataset purposes.

Contribution 2: Scalable iterative rejection sampling supervised finetuning approach

Description: The authors introduce a training methodology using iterative rejection sampling with supervised finetuning that provides stable and efficient training at scale. This semi-online approach avoids distribution shift issues while remaining computationally tractable compared to full online RL methods.

This contribution was assessed against **4 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Statistical rejection sampling improves preference optimization

URL: View paper

Brief Assessment

Statistical Rejection Sampling[44] focuses on preference optimization using rejection sampling to construct preference pairs from an estimated optimal policy, not on training evaluators. The candidate addresses a fundamentally different problem domain (preference learning for generation) rather than evaluator training with iterative rejection sampling supervised finetuning.

---

2. Aryabhata: An exam-focused language model for JEE Math

URL: View paper

Brief Assessment

Aryabhata[46] uses rejection sampling for SFT on a math reasoning model, but focuses on exam-specific curriculum learning rather than the multi-task evaluator training methodology described in the original paper. The candidate does not demonstrate prior work on the original's semi-online RS-SFT approach for training evaluators at scale.

---

3. STARS: Segment-level Token Alignment with Rejection Sampling in Large Language Models

URL: View paper

Brief Assessment

STARS[47] focuses on inference-time alignment using segment-level rejection sampling for decoding, not on training evaluators. The original paper describes a training methodology for evaluators using iterative rejection sampling with supervised finetuning, while STARS[47] applies rejection sampling during generation/inference without model weight updates.

---

4. GMAI-VL-R1: Harnessing Reinforcement Learning for Multimodal Medical Reasoning

URL: View paper

Brief Assessment

GMAI-VL-R1[45] focuses on multimodal medical reasoning with RL enhancement, using rejection sampling for medical data synthesis. The original paper addresses general evaluator training across multiple domains with a semi-online RS-SFT approach to avoid distribution shift at scale, which is a different application context and technical focus.

Contribution 3: FARE family of foundational automatic reasoning evaluators

Description: The authors develop FARE, a family of 8B and 20B parameter evaluators trained on their multi-task dataset. These models are evaluated on static benchmarks and real-world applications including inference-time reranking, verification during RL training, and domain adaptation.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Agent0-VL: Exploring Self-Evolving Agent for Tool-Integrated Vision-Language Reasoning

URL: View paper

Brief Assessment

Agent0-VL[48] focuses on self-evolving vision-language agents with tool-integrated reasoning for multimodal tasks, not on training foundational automatic reasoning evaluators for text-based reasoning evaluation across multiple tasks and domains.

---

2. Towards Continuous Intelligence Growth: Self-Training, Continual Learning, and Dual-Scale Memory in SuperIntelliAgent

URL: View paper

Brief Assessment

SuperIntelliAgent[52] focuses on a self-training framework with a diffusion model learner and LLM verifier for continual learning, not on developing foundational automatic reasoning evaluators for multi-task evaluation across diverse domains.

---

3. SATQuest: A Verifier for Logical Reasoning Evaluation and Reinforcement Fine-Tuning of LLMs

URL: View paper

Brief Assessment

SATQuest[51] focuses on logical reasoning evaluation using SAT-based problems (CNF instances), not on training general-purpose automatic reasoning evaluators for diverse tasks like pairwise comparison, step-level evaluation, or verification across multiple domains.

---

4. AORO: Auto-Optimizing Reasoning Order for Multi-Hop Question Answering

URL: View paper

Brief Assessment

AORO[50] focuses on optimizing the reasoning order for multi-hop question answering retrieval, not on developing foundational automatic evaluators for reranking, verification, or continual finetuning in reasoning-centric domains.

---

5. Continuous Automated Model EvaluatiOn (CAMEO)-Perspectives on the future of fully automated evaluation of structure prediction methods.

URL: View paper

Brief Assessment

CAMEO[53] focuses on automated evaluation of protein structure prediction methods in computational biology, not on training foundational automatic reasoning evaluators for language models. The domains and methodologies are entirely different.

---

6. Utilizing large language models for question answering in task-oriented dialogues

URL: View paper

Brief Assessment

Task-Oriented Question Answering[49] focuses on question answering in task-oriented dialogues using large language models, not on developing foundational automatic reasoning evaluators for reranking, verification, or continual finetuning in reasoning-centric domains.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Foundational Automatic Evaluators: Scaling Multi-Task Generative Evaluator Training for Reasoning-Centric Domains View paper
- [1] Praetor: A Fine-Grained Generative LLM Evaluator with Instance-Level Customizable Evaluation Criteria View paper
- [2] Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models View paper
- [3] Towards responsible development of generative AI for education: An evaluation-driven approach View paper
- [4] TAIA: Telco Generative AI-powered Multi-Agent Assistant for managing Cloud-Native Networks View paper
- [5] Diffusion Model is an Effective Planner and Data Synthesizer for Multi-Task Reinforcement Learning View paper
- [6] Direct evaluation of chain-of-thought in multi-hop reasoning with knowledge graphs View paper
- [7] Towards Unified Multi-Modal Personalization: Large Vision-Language Models for Generative Recommendation and Beyond View paper
- [8] Beyond Task-Specific Reasoning: A Unified Conditional Generative Framework for Abstract Visual Reasoning View paper
- [9] J4R: Learning to Judge with Equivalent Initial State Group Relative Policy Optimization View paper
- [10] Wireless multi-agent generative AI: From connected intelligence to collective intelligence View paper

- [11] BeyondBench: Benchmark-Free Evaluation of Reasoning in Language Models View paper
- [12] UniEvent: Unified Generative Model with Multi-Dimensional Prefix for Zero-Shot Event-Relational Reasoning View paper
- [13] Enhancing explainable recommendations: Integrating reason generation and rating prediction through multi-task learning View paper
- [14] MechAgents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge View paper
- [15] Multi-Agent Generative AI: Coordinated Synthesis for Complex Problem-Solving View paper
- [16] Generative Evaluation of Complex Reasoning in Large Language Models View paper
- [17] MTD-GPT: A Multi-Task Decision-Making GPT Model for Autonomous Driving at Unsignalized Intersections View paper
- [18] Flex-Judge: Text-Only Reasoning Unleashes Zero-Shot Multimodal Evaluators View paper
- [19] ReFeR: Improving Evaluation and Reasoning through Hierarchy of Models View paper
- [20] CognitiveOS: Large Multimodal Model Based System to Endow Any Type of Robot with Generative AI View paper
- [21] SAI4EO: Generative AI for Earth Observation Tasks View paper
- [22] ARM-Thinker: Reinforcing Multimodal Generative Reward Models with Agentic Tool Use and Visual Reasoning View paper
- [23] Enhancing Large Language Models with Neurosymbolic Reasoning for Multilingual Tasks View paper
- [24] ReasonIR: Training Retrievers for Reasoning Tasks View paper
- [25] A Generative AI-Enhanced Case-Based Reasoning Method for Risk Assessment: Ontology Modeling and Similarity Calculation Framework View paper
- [26] MMGR: Multi-Modal Generative Reasoning View paper
- [27] ProfVLM: A Lightweight Video-Language Model for Multi-View Proficiency Estimation View paper
- [28] Scaling Evaluation-time Compute with Reasoning Models as Process Evaluators View paper
- [29] Validating Computational Deliberation: An Empirical Analysis of Governed vs. Generative AI Reasoning View paper
- [30] A Modular Multitask Reasoning Framework Integrating Spatio-temporal Models and LLMs View paper
- [31] Automating Expert-Level Medical Reasoning Evaluation of Large Language Models View paper
- [32] Prompting Contrastive Explanations for Commonsense Reasoning Tasks View paper
- [33] Autonomous Evaluation of LLMs for Truth Maintenance and Reasoning Tasks View paper
- [34] FullDiT: Multi-Task Video Generative Foundation Model with Full Attention View paper
- [35] Chain-of-Thought Reasoning Evaluation Framework for Question Answering System View paper
- [36] SocREval: Large Language Models with the Socratic Method for Reference-Free Reasoning Evaluation View paper
- [37] Multi-Task Training with In-Domain Language Models for Diagnostic Reasoning View paper
- [38] Multi-Task Learning of Japanese How-to Tip Machine Reading Comprehension by a Generative Model View paper
- [39] Learning to Solve Abstract Reasoning Problems with Neurosymbolic Program Synthesis and Task Generation View paper
- [40] Exploration for Multi-task Reinforcement Learning with Deep Generative Models View paper
- [41] Abstract Reasoning via Logic-guided Generation View paper
- [42] CS-NLP team at SemEval-2020 Task 4: Evaluation of State-of-the-art NLP Deep Learning Architectures on Commonsense Reasoning Task View paper
- [43] Intelligent Generative Design: A New Mechanical Design Concept View paper
- [44] Statistical rejection sampling improves preference optimization View paper
- [45] GMAI-VL-R1: Harnessing Reinforcement Learning for Multimodal Medical Reasoning View paper
- [46] Aryabhata: An exam-focused language model for JEE Math View paper
- [47] STARS: Segment-level Token Alignment with Rejection Sampling in Large Language Models View paper
- [48] Agent0-VL: Exploring Self-Evolving Agent for Tool-Integrated Vision-Language Reasoning View paper
- [49] Utilizing large language models for question answering in task-oriented dialogues View paper
- [50] AORO: Auto-Optimizing Reasoning Order for Multi-Hop Question Answering View paper
- [51] SATQuest: A Verifier for Logical Reasoning Evaluation and Reinforcement Fine-Tuning of LLMs View paper
- [52] Towards Continuous Intelligence Growth: Self-Training, Continual Learning, and Dual-Scale Memory in SuperIntelliAgent View paper
- [53] Continuous Automated Model EvaluatiOn (CAMEO)-Perspectives on the future of fully automated evaluation of structure prediction methods. View paper
- [54] TerraGen: A Unified Multi-Task Layout Generation Framework for Remote Sensing Data Augmentation View paper
- [55] Enhancing logical reasoning in large language models through graph-based synthetic data View paper
- [56] SHIFT: A Synthetic Driving Dataset for Continuous Multi-Task Domain Adaptation View paper
- [57] Promptonomyvit: Multi-task prompt learning improves video transformers using synthetic scene data View paper
- [58] Versaprm: Multi-domain process reward model via synthetic reasoning data View paper
- [59] Towards Hierarchical Multi-Step Reward Models for Enhanced Reasoning in Large Language Models View paper
- [60] Enhancing Domain-Specific Retrieval-Augmented Generation: Synthetic Data Generation and Evaluation using Reasoning Models View paper
- [61] Internbootcamp technical report: Boosting llm reasoning with verifiable task scaling View paper
- [62] Beyond Intelligence: The Synthetic Cognitive Augmentation Network Using Experts View paper
- [63] SMIR: Efficient Synthetic Data Pipeline To Improve Multi-Image Reasoning View paper