

ICML '24 notes and interesting posters

April 7, 2025

Montag Vormittag

Predictive attribution

- loss → what if one element changed? Leave One Out (LOO)
- can be solved in closed form for Linear Regression, Logistic Regressions, NNs,..

NN operator learning

- PDE learning → nonlinearities keys
- optimize these jointly with NN
- rewatch talk! (Physics of LLMs too)

Strategic Learning & behaviour

- behavior (human) influences ML decisions and vice-versa
- classifiers: transparent or opaque (e.g. Schufa? what are the merits)
- includes the social burden of ML system in loss (never explained how??)
- people move in classifier feature space - towards better decision rule, independent from position!
 ⇒ classifiers create demand!
- strategic modification ⇒ strategic participation in system!
 - is it worth to participate at all? - your fairness might be skewed!
 - e.g.: people might not apply in the first chance, showing a fair selection bias to "pre-selection"
 - fairness is opaque!
- possible solution: causality of change?

Montag Nachmittag

Distribution-Free UCQ

- problem of conformal splits: variability: $\mathcal{O}(\sqrt{n})$
- no split conformal prediction
 - problem when scoring on training points → $\not\perp$ overfitting, all $s(x) = 0$
 - go back to classical CP - leave-one-out train for all → $\not\perp \mathcal{O}(n^2)$
 - jackknife+ → problem if node unstable
- adaptability of CP using running adaption γ as basis (AgACI)

GNNs

- node-level tasks:
 - node embeddings (optimized based on similarity+random walks)
 - Problems: incorporate structure, adding data
- GNN - aggregate information ~ CNN
- message passing to neighbors
- final output layer: based on task!
- GCN: passing adjacency and diagonal matrix iteratively
- GraphSage
- GAT: Graph Transformer: attend on neighbors
- instead of future predict node
 - problem: convey position to transformer (usually sine embeddings)
 - output: node embeddings
 - regular sinusoidal embeddings with graph laplacian + learnable embeddings
 - problem : $O(n^2)$
- GraphGPS: message passing + transformer

Dienstag Vormittag

Unapologetic Openness

- why openness? → ecosystem ↑, community thrives!
- not just philanthropy
- why not: time advantage, could be used in harmful ways
- LLM Open Source: human feedback ↳ meta wants end user feedback, but is missing that for training...

Genie

- train Video model with action tokens for 16 frames
- goal: agents can use and understand sim

Arrows in time

- forward/backward CE of LLM
- Related to language/information theorem of Shannon
- Forward pass has a lower loss - indicates an arrow in time!
- Across all languages!
- gap increases with model size, across multiple model types
- origin: primes $p_1, p_2, p_1 \times p_2 = n$ - multiplication easy, factorisation is not
- causality?, very data-intense, not clear if it applies to other data

Transformers for pretraining Universal Forecasts: MOIRAI

- challenges: cross-frequency
- patch-based forecasting +masked
- multivariate: flattened, different encoding
- Future Work: combine with text?

Potential of Transformers for Timeseries prediction: SAMFormer

- robustness against time shift
- custom training routine: SAM
- very simple, better than MOIRAI-zero-Shot
- The same architecture works well for many systems

Mittwoch Vormittag

African Language Datasets

- translations missing, important to bring policy decisions to citizens
- no clear text available - only as PDFs or similar, only 10 % is translated!
- alignment issues
- voice dataset being built
- translate scientific content at scale
- code mixing problems (NLP)
- Lelapa (home): communicating in African languages
 - community, from scratch: 45% women!
 - legal aspects of AI largely unknown, a lot of workshops

Position: Measure Diversity, don't just claim it!

- collect geographically diverse dataset, diversity definition matters - which level of diversity, ...
- diversity can never be objective → values encode information (e.g. political)
- measurement still fundamental for ML
- measurement theory (social science), e.g. socioeconomic status based on many factors, only indirect measure possible
 - conceptualize
 - operationalize
 - evaluate
 - ?
- → scale ≠ diversity ≠ unbiased
- not much quality reported
- evaluation usually only on newer models
- measure diversity *within* dataset → problem: level of diversity, unknown definition!

Mittwoch Nachmittag

SceneCraft: Text2Scene

- challenge: semantic relationship not controllable
- solution: LLM agents repeat generative approach+function generation to build skills automatically
 1. asset list → CLIP search for similar assets
 2. scene decomposition using LLM
 3. layout checked for each object → semantics/relationships!
 4. critique & adopt functions
- extended to movie generation → movie poet, a bit weak

ChatGPT moderation at scale

- downsides to ChatGPT: learning hindered, factually incorrect
- indicator adjectives show that GPT use is on the rise
- indistinguishable from human?
- corpus-level detection (percentage)
- ~10% to 17% usage, Nature almost 0!
- Multimodal α estimation using known distributions
- ground truth generated by LLM generated reviews for papers before 2020, temporal split!
- modeling TF of on adjectives for probabilities
- common GPT detectors worse!
- BERT-based detectors weak
- deadline effect: more usage!
- more replies: less usage (more involvement!)
- only works globally, not necessarily bad - can be used as an indicator, not individual blame!

Stealing part of a production LLM

- finding single values of LLM responses
- singular value decomposition: after a certain number of stops steep falloff of values - indicates the limit of the last layer!
- indicates output subspace - consequently, last layer size!
- final layers can be learned too:
$$Q = U\Sigma V^T \tag{1}$$
- can be learned using SVD
- is worth stealing, as ML can be used to generate profit now!

MagicLens: Self-Supervised Image Retrieval

- usually in image retrieval: most *identical* image
- here: guide image + search intent - retrieve semantically relevant image!
- problem: training data:
 - websites with 2+ images as adjacent images, with nearby text
 - filter out ads (Google cannot disable their ads??)
- contrastive loss, good results
- outperforms SOTA image retrieval
- extremely good semantic retrieval

Donnerstag Vormittag

Position: Opportunities exist for ML+Fusion

- high energy output, tritium production, economics
- disruption prediction
- simulation & dynamics modeling - physics are incomplete!
- partial observability (related to our HO problem)
- controls problems, experiment design
- material design

HEPT: High Energy Particle Transformer

- Particle cloud embeddings for transformers

Donnerstag Nachmittag

Uncertainties for LLM

- perturb inputs instead of ensemble LLM
- disentangle → epistemic/aleatoric
- prompting/finetuning diversity

AlphaFlow Meets Flow Model Matching

- distribution of structures in protein folding
- generative modeling!
- AlphaFlow denoises 3D structure from template + protein

Freitag

ML4ESM: Towards improved cloud modelling

ML4ESM: Climate Set

- Climate models: future emissions → how does the climate react?
- Multiple socio-economic pathways
- ~ 390 days/simulation!
- problem: resolution scales $O(r^3)$
- ML: can help downsampling, parametrization, *emulation*
- Problems: distribution shift, data-based, high uncertainty in models (5 K)

ML4ESM: ML and Climate Change

- ML not problem/application driven!
- problem: limited resources, sparsely labeled data
- domain knowledge required - reduces compute significantly!
- Climate Simulation
 - reduce the resolution of simulation, scale up using super-resolution
 - keep physical constraints in mind
 - mapping to continuous functions: related to neural operator learning

ML4ESM: PDE+phys. Constraints+Spectral

ML4ESM: DDPM: Deep Denoising Physical Models

- PDE model using diffusion process → enables uncertainty modeling!
- constraint diffusion process!

Samstag

GRaM: Platonic Representation Hypothesis

- models learn same “representation”
- converges to same clues in feature spaces (e.g. dogs detector to ears, ...)
- “Rosetta neurons” - same representation across many models → is there convergence?
 - H1: different representation
 - H2: or same representation? (good models \Leftrightarrow similar representation)
- Language+Visualisation: do models converge - some indications:
 - Use kernel to map similarity between models, map different concepts of e.g. GPT, ImageNet
 - result: language represents similar concepts as vision!
 - a lot of limitations, currently only 0.2/1, does not converge to reality

Sociotechnical Evaluation of AI

- layers: capabilities, human interactions, systemic impacts
- problem: only technical aspects of AI considered & mostly textual evaluation
- e.g. textual evaluation:
 - replica users, mental health impact
 - stackoverflow activity drop after ChatGPT release
 - homogenization of creative writing: least creative get uplift, most creative reduce creativity - narrowing of the spectrum!
- studies: synthetic simulation?

AI safety institute (UK)

- evaluation of AI: misuse, societal impacts (long term!), autonomous systems (loss of control, safeguards for agents and tools!)

Future of video generation - beyond data and scale

- currently: imperfect control over semantics
- research: single video model, instead of foundational model → can be used to split background-/foreground, alpha & recombine

Adversarial Perturbations cannot Reliably protect artists from generative AI

- existing adversarial perturbation can easily be bypassed using:
 - Gaussian Filters
 - One Diffusion step
 - ...

CopyCat

- Remove copyrighted characters
- Using: negative prompting (post hoc - open models can easily circumvent that!)

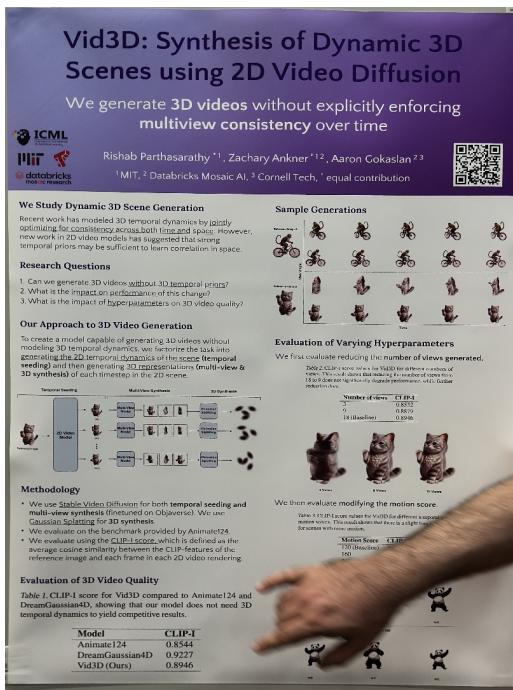
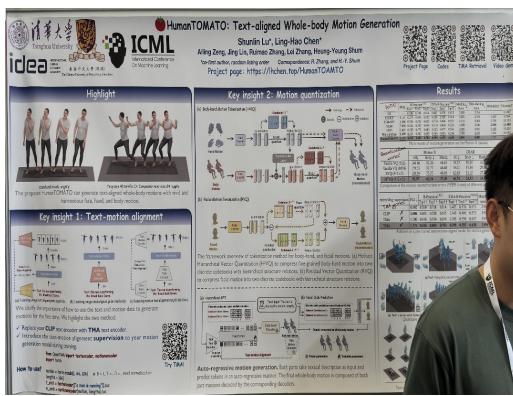
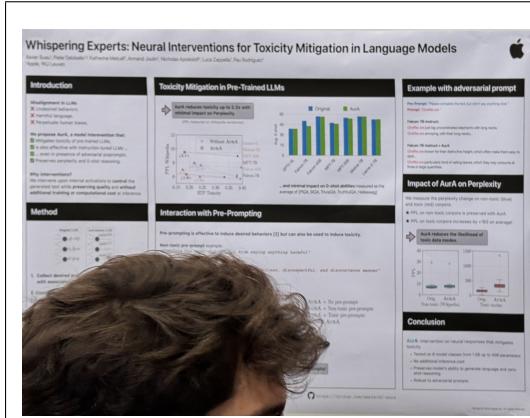
Posters

Table 1:

Poster	Information																																
<p>Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text Abhimanyu Hars - Avi Schwarzschild - Valeria Cheprenova - Hamid Kazemi - Aniruddha Saha - Mican Geiping - Tom Goldstein</p> <ul style="list-style-type: none"> Binoculars achieves state-of-the-art zero-shot accuracy. Binoculars is capable of spotting machine text from a range of modern LLMs without any model-specific modifications. Binoculars detects over 90% of generated samples from ChatGPT at a false positive rate of 0.01%. <p>The error of an AI detector needs to be examined carefully!</p> <p>Loss of model M_i on string s</p> $B_{M_1, M_2}(s) = \frac{\log \text{PPL}_{M_1}(s)}{\log x\text{PPL}_{M_1, M_2}(s)}$ <p>Loss of model M_i using each next token predictor from M_j as the original</p> <p>$\log \text{PPL}_M(s) = -\frac{1}{T} \sum_{t=1}^T \log(Y_{it})$</p> <p>$\log x\text{PPL}_{M_1, M_2}(s) = -\frac{1}{T} \sum_{t=1}^T M_1(s_t) \cdot \log(M_2(s_t))$</p>	<p>Block-level Text Spotting with LLMs <i>Ganesh Bannur, Bharadwaj Amrutar</i> Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text http://arxiv.org/abs/2406.13208v1</p>																																
<p>Scaling Rectified Flow Transformers for High-Resolution Image Synthesis Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, Robin Rombach</p> <p>Scaling Image Synthesis</p> <p>Learning ODEs Between Distributions</p> <p>Finding the Best Architecture for Multimodal Training</p> <p>Learning the Important Things</p> <p>Validation Loss is a Strong Predictor of Performance</p> <p>Performance Evaluation</p>	<p>Scaling Rectified Flow Transformers for High-Resolution Image Synthesis <i>Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, Robin Rombach</i> Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, Robin Rombach http://arxiv.org/abs/2403.03206v1</p>																																
<p>SCENE-Net V2: Interpretable Multiclass 3D Scene Understanding with Geometric Priors Diogo Lavado, Cláudia Soares and Alessandra Micheletti</p> <p>Introduction</p> <p>GENEOs</p> <p>To build geometric priors, we leverage Group Equivariant Non-expansive Operators (GENEOs) [1]. These operators provide a measure of the world, based on geometric properties of convolutional kernels. Unlike convolutional kernels, our GENEOs are not tied to the underlying geometry of 3D scenes; they are parameterized with meaningful features.</p> <p>SCENE-Net V2 is a gray-box model that pairs geometric interpretability and general application</p> <p>White-box feature extraction phase with 540 meaningful shape parameters</p> <p>In the GENEo Layer, we instantiate m GENEo-kernels from m families of geometric priors. Such families are defined by meaningful shape parameters, such as the radius of a cylinder. They are then combined into n observations through convex combinations, creating more complex feature extraction outputs.</p> <p>A CNN-based feature extraction process with an analogous architecture contains 21.4K parameters.</p> <p>Experiments</p> <p>The Performance of SCENE-Net V2</p> <table border="1"> <thead> <tr> <th>Method</th> <th>Model</th> <th>#Parameters</th> <th>Top-1</th> </tr> </thead> <tbody> <tr> <td>PointNet [Qi et al., 2017a]</td> <td>None</td> <td>~100K</td> <td>49.5%</td> </tr> <tr> <td>KPConv [Thomé et al., 2019]</td> <td>None</td> <td>~100K</td> <td>50.5%</td> </tr> <tr> <td>PIConv [Wu et al., 2021]</td> <td>None</td> <td>~100K</td> <td>51.2%</td> </tr> <tr> <td>Point Transformer V1 [Wu et al., 2021a]</td> <td>None</td> <td>~100K</td> <td>51.5%</td> </tr> <tr> <td>Point Transformer V2 [Wu et al., 2021b]</td> <td>None</td> <td>~100K</td> <td>51.8%</td> </tr> <tr> <td>SCENE-Net V2 (Ours)</td> <td>None</td> <td>~240K</td> <td>52.2%</td> </tr> <tr> <td>SCENE-Net V2 + CNN (Ours)</td> <td>CNN</td> <td>~240K</td> <td>52.2%</td> </tr> </tbody> </table> <p>Future Work</p> <ul style="list-style-type: none"> Using SCENE-Net V2 as a feature extraction tool for SOTA benchmarks. Applying GENEos directly onto raw 3D point clouds. 	Method	Model	#Parameters	Top-1	PointNet [Qi et al., 2017a]	None	~100K	49.5%	KPConv [Thomé et al., 2019]	None	~100K	50.5%	PIConv [Wu et al., 2021]	None	~100K	51.2%	Point Transformer V1 [Wu et al., 2021a]	None	~100K	51.5%	Point Transformer V2 [Wu et al., 2021b]	None	~100K	51.8%	SCENE-Net V2 (Ours)	None	~240K	52.2%	SCENE-Net V2 + CNN (Ours)	CNN	~240K	52.2%	<p>SCENE-Net V2 is a gray-box model that pairs geometric interpretability and (1) Bergomi, M. G., Frosini, P., Giorgi, D., and Quercioli, N. Towards a topolog</p> <p>SCENE-Net V2 is a gray-box model that pairs geometric interpretability and</p>
Method	Model	#Parameters	Top-1																														
PointNet [Qi et al., 2017a]	None	~100K	49.5%																														
KPConv [Thomé et al., 2019]	None	~100K	50.5%																														
PIConv [Wu et al., 2021]	None	~100K	51.2%																														
Point Transformer V1 [Wu et al., 2021a]	None	~100K	51.5%																														
Point Transformer V2 [Wu et al., 2021b]	None	~100K	51.8%																														
SCENE-Net V2 (Ours)	None	~240K	52.2%																														
SCENE-Net V2 + CNN (Ours)	CNN	~240K	52.2%																														

Continued on next page

Table 1: (Continued)



Whispering Experts: Neural Interventions for Toxicity Mitigation in Language Models *average of (P^aOA , $SIOA$, $TriviaQA$, $TruthfulGA$, $Helwasql$)*

Whispering Experts: Neural Interventions for Toxicity Mitigation in Language Models

Spinning Down a Black Hole With Scalar Fields *Chris M. Chambers, William A. Hiscock, Brett Taylor*

HumanTOMATO: Text-aligned Whole-body Motion Generation

<http://dx.doi.org/10.1103/PhysRevLett.78.3249>

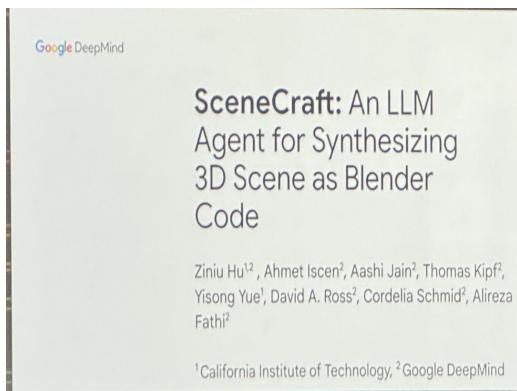
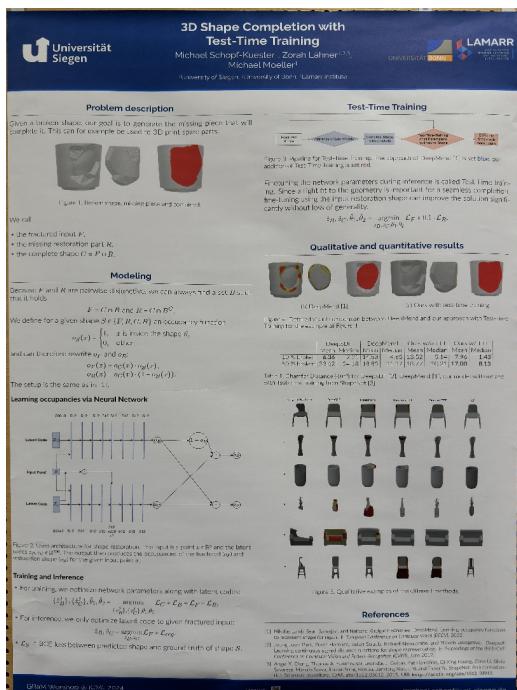
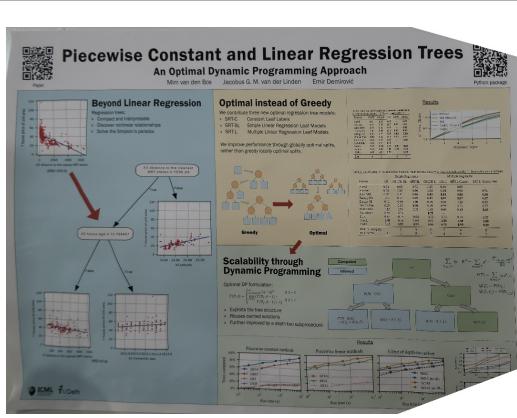
The effects of Gribov copies in 2D gauge theories *D. Dudal, S. P. Sorella, N. Vandersickel, H. Verschelde*

Vid3D: Synthesis of Dynamic 3D

<http://dx.doi.org/10.1016/j.physletb.2009.08.055>

Continued on next page

Table 1: (Continued)



Efficient Regularized Piecewise-Linear Regression Trees Leonidas Lefakis, Oleksandr Zadorozhnyi, Gilles Blanchard

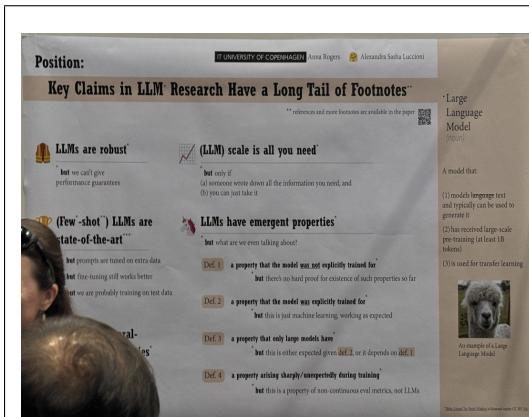
Piecewise Constant and Linear Regression Trees
<http://arxiv.org/abs/1907.00275v1>

Refusion: Enabling Large-Size Realistic Image Restoration with Latent-Space Diffusion Models Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, Thomas B. Schön
fine-tuning using the input restoration shape can improve the solution signifi-
<http://arxiv.org/abs/2304.08291v1>

The HulC: Confidence Regions from Convex Hulls Arun Kumar Kuchibhotla, Sivaraman Balakrishnan, Larry Wasserman
Agent for Synthesizing
<http://arxiv.org/abs/2105.14577v2>

Continued on next page

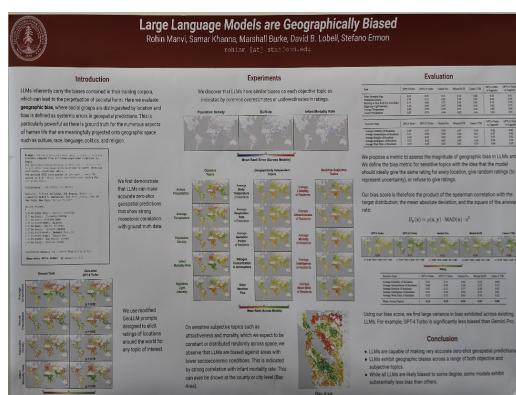
Table 1: (Continued)



Position: Key Claims in LLM Research Have a Long Tail of Footnotes *Anna Rogers, Alexandra Sasha Luccioni*

Key Claims in LLM Research Have a Long Tail of Footnotes*

<http://arxiv.org/abs/2308.07120v2>

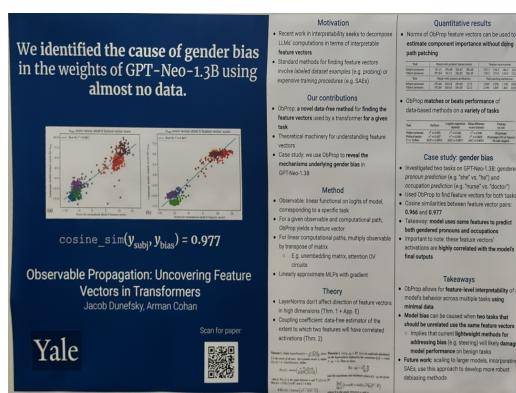


Large Language Models are Geographically Biased

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, Stefano Ermon

Large Language Models are Geographically Biased

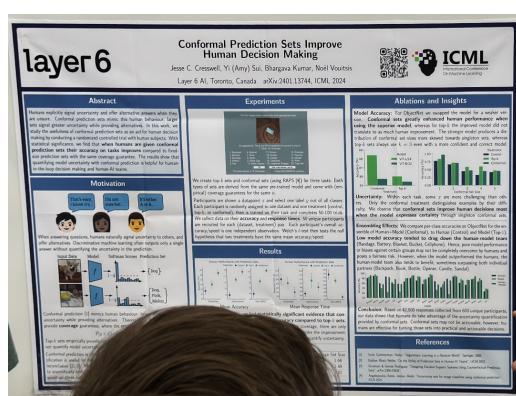
<http://arxiv.org/abs/2402.02680v1>



MISGENDERED: Limits of Large Language Models in Understanding Pronouns *Tamanna Hossain, Sunipa Dev, Sameer Singh*

in the weights of GPT-Neo-1.3B using
<http://arxiv.org/abs/2306.03950v2>

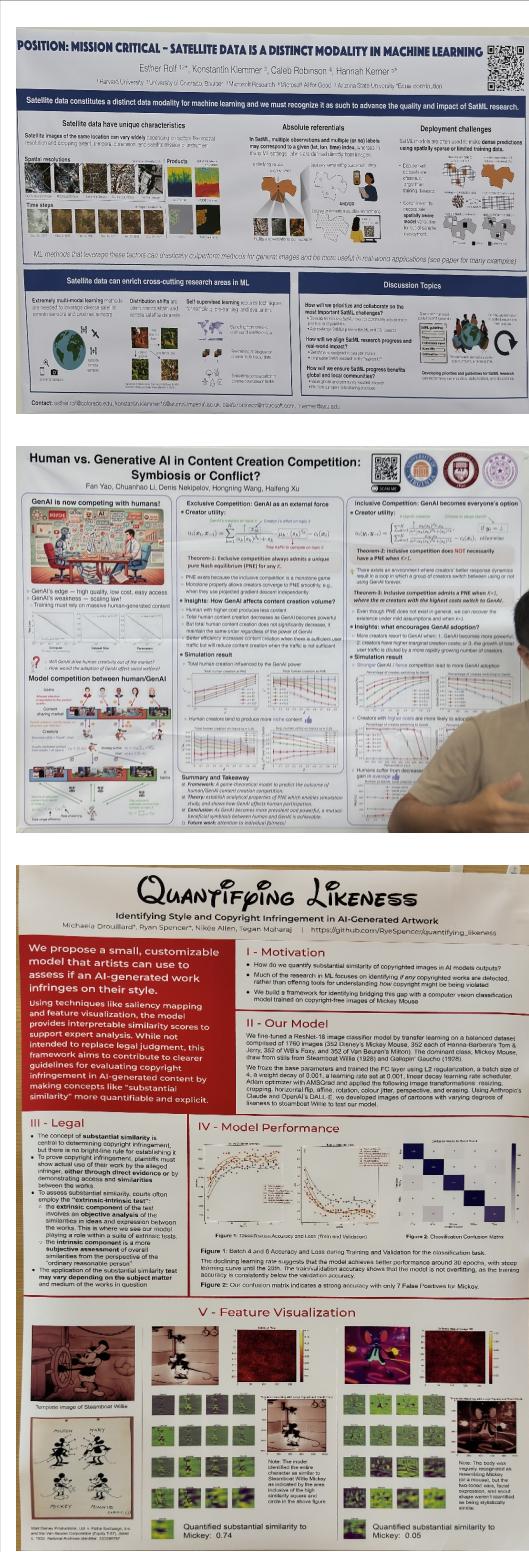
1 *W* **2** *W* **3** *W* **4** *W* **5** *W* **6** *W* **7** *W* **8** *W* **9** *W* **10** *W* **11** *W* **12** *W* **13** *W* **14** *W* **15** *W* **16** *W* **17** *W* **18** *W* **19** *W* **20** *W* **21** *W* **22** *W* **23** *W* **24** *W* **25** *W* **26** *W* **27** *W* **28** *W* **29** *W* **30** *W* **31** *W* **32** *W* **33** *W* **34** *W* **35** *W* **36** *W* **37** *W* **38** *W* **39** *W* **40** *W* **41** *W* **42** *W* **43** *W* **44** *W* **45** *W* **46** *W* **47** *W* **48** *W* **49** *W* **50** *W* **51** *W* **52** *W* **53** *W* **54** *W* **55** *W* **56** *W* **57** *W* **58** *W* **59** *W* **60** *W* **61** *W* **62** *W* **63** *W* **64** *W* **65** *W* **66** *W* **67** *W* **68** *W* **69** *W* **70** *W* **71** *W* **72** *W* **73** *W* **74** *W* **75** *W* **76** *W* **77** *W* **78** *W* **79** *W* **80** *W* **81** *W* **82** *W* **83** *W* **84** *W* **85** *W* **86** *W* **87** *W* **88** *W* **89** *W* **90** *W* **91** *W* **92** *W* **93** *W* **94** *W* **95** *W* **96** *W* **97** *W* **98** *W* **99** *W* **100** *W*



layer 6 partners (*Backpack, Book, Bottle, Opener, Candle, Sandal*).
layer 6

Continued on next page

Table 1: (Continued)



Missing-modality Enabled Multi-modal Fusion Architecture for Medical Data *Muyu Wang, Shiyu Fan, Yichen Li, Hui Chen*

POSITION: MISSION CRITICAL - SATELLITE DATA IS A DISTINCT MODALITY IN MACHINE LEARNING

<http://arxiv.org/abs/2309.15529v1>

Human vs. Generative AI in Content Creation Competition: Symbiosis or Conflict?

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, Haifeng Xu

Human vs. Generative AI in Content Creation Competition:

<http://arxiv.org/abs/2402.15467v1>

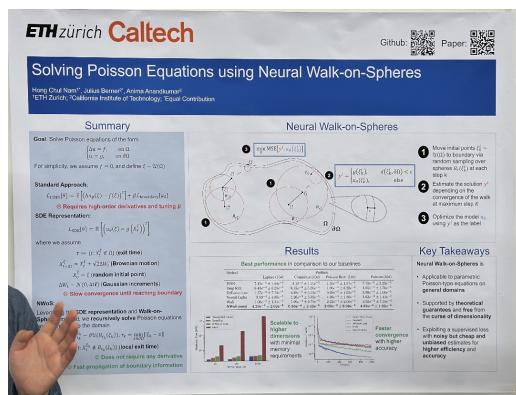
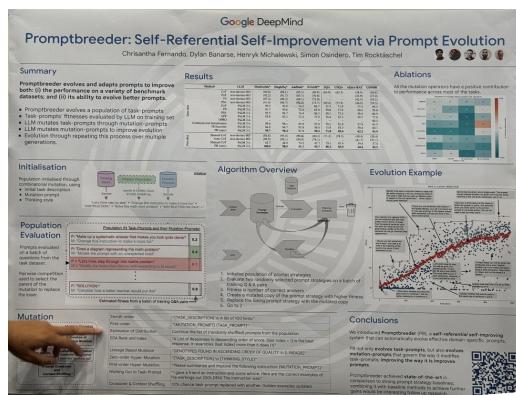
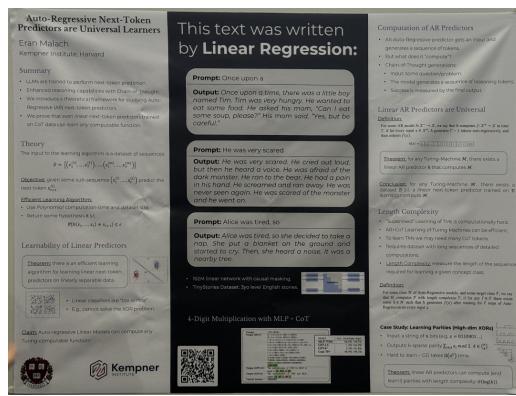
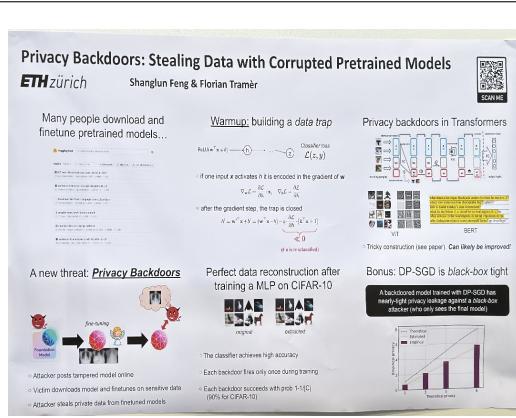
Multimodal Crop Type Classification Fusing Multi-Spectral Satellite Time Series with Farmers Crop Rotations and Local Crop Distribution *Valentin Barriere, Martin Claverie*

QUANTiFpiNG LiKENESS

<http://arxiv.org/abs/2208.10838v1>

Continued on next page

Table 1: (Continued)



Privacy Backdoors: Stealing Data with Corrupted Pretrained Models Shanglun Feng, Florian Tramèr

Privacy Backdoors: Stealing Data with Corrupted Pretrained Models
<http://arxiv.org/abs/2404.00473v1>

by Linear Regression: Objective: given some subsequence x, \dots, x_r , predict the next token x_{r+1} by Linear Regression:

Promptbreeder: Self-Referential Self-Improvement Via Prompt Evolution Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, Tim Rocktäschel

Promptbreeder: Self-Referential Self-Improvement via Prompt Evolution
<http://arxiv.org/abs/2309.16797v1>

Integral Equation Approach to Stationary Stochastic Counting Process with Independent Increments Enzhi Li

Solving Poisson Equations using Neural Walk-on-Spheres
<http://arxiv.org/abs/1811.07262v1>

Continued on next page

Table 1: (Continued)

Motivation: Users Have Diverse Search Intents
Query Image → Search Intents → MagicLens → Image

Solution: Mining Naturally Occurring Image Pairs → Query Image / Input Image → Model → Top 1000 Results

Modeling: Simple Contrastive & Parameter-Sharing

Encoder: Vision Encoder + Language Encoder

Attention Pooling: Self Attention + K, V, E

Final Output: Scored and sorted results

Conclusion: MagicLens models are 1) Unsupervised, 2) Low-latency, 3) Unified, and 4) Open-ended.

MagicLens: Self-Supervised Image Retrieval with Open-Ended Instructions Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhui Chen, Yu Su, Ming-Wei Chang

MagicLens: Next-Generation Image Retrieval Models

<http://arxiv.org/abs/2403.19651v2>

Introduction: How to computationally assess how interpretable a model is? This addresses this using computational theory – the harder it is to generate an explanation, the less interpretable a model is.

Local Sufficient Reason: $f(\mathbf{x}) = f(\mathbf{x}_1, \mathbf{x}_2) = f(\mathbf{x})$

Global Sufficient Reason: $f(\mathbf{x}) = f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = f(\mathbf{x})$

Explanation forms we analyze:

- Feature selection – selecting which local vs. global suffices reasons
- Identifying local vs. global redundant or necessary features
- A probability notion of sufficiency – whether local or global

Conclusion:

- We prove that there is often a strict complexity gap between obtaining an explanation locally vs. globally.
- In some cases, such gaps justify folklore claims (e.g., linear models can be interpreted locally using their weights but not globally).
- In other cases, they yield unexpected outcomes (e.g., neural networks and decision trees are easier to interpret globally than locally).

Shahaf Bassan, Guy Amir, Guy Katz; The Hebrew University of Jerusalem

Local vs. Global Interpretability: A Computational Complexity Perspective Shahaf Bassan, Guy Amir, Guy Katz

Local vs. Global Interpretability: A Computational Complexity Perspective

<http://arxiv.org/abs/2406.02981v2>

Scaling down deep learning with MNIST-1D

Sam Grysman^{1,2}
Dmitry Kehak^{3,4}

¹Oregon State University, USA
²TUM, Germany
³University of Cologne, Germany
⁴Haberdashers' Aske's Boys' School, UK

From MNIST to MNIST-1D

MNIST-1D allows to study deep learning phenomena on a laptop

Benign overfitting: Shows that overfitting is not necessarily bad for generalization.

Double descent: Shows that double descent is not always monotonic.

Guillotine regularization in self-supervised learning: Shows that guillotine regularization can help in self-supervised learning.

Meta-learning: Shows the relationship between learned meta-weights and learned activation functions.

deep learning nn. Convid(25, 25, 3, stride=2, padding=1), deep learning

Abstract: This paper introduces mean-field chaos diffusion models (MCCMs), a new class of noise-based generative models trained for mean-field particle systems. MCCMs address the curse of dimensionality in mean-field systems. We propose a novel type of score matching objective for objects with inherent chaoticity.

Proposed Method: Based on the developed variational analysis, we propose a new type of score matching objective for objects with inherent chaoticity.

Theorem 4.3: If \mathcal{L} is convex, the proposed MCFMs achieve the augmented chaotic energy minimization problem: $\min_{\theta} \mathcal{L}(\theta) + \lambda \mathbb{E}_{\mathbf{x} \sim p_{\theta}} [\mathcal{L}(f_{\theta}(\mathbf{x})) - \mathbb{E} \log p_{\theta}(\mathbf{x})]$.

Empirical Results: Synthetic Dataset, Variational Analysis, Real-world Dataset (ShapeNet, Mesh2Shape), 3D Shape Generation of Sparse LiDAR Data, Density of Particle Branching Process, Performance Comparison with SOTA 3D Shape Models.

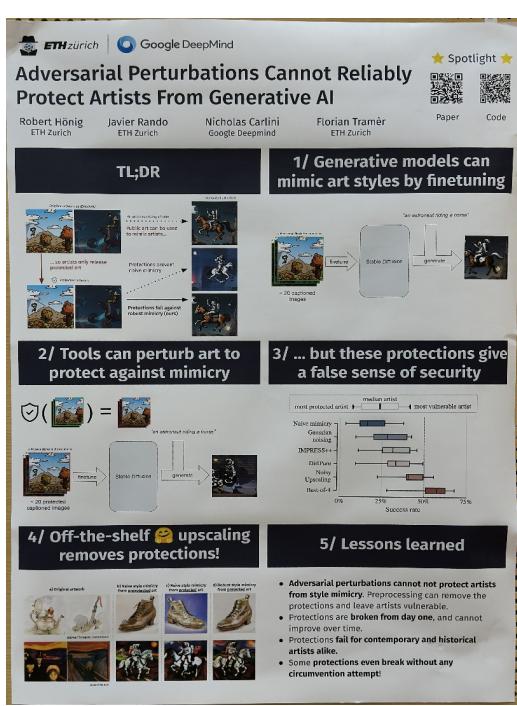
Active matter beyond mean-field: Ring-kinetic theory for self-propelled particles Yen-Liang Chou, Thomas Ihle

Mean-field Chaos Diffusion Models

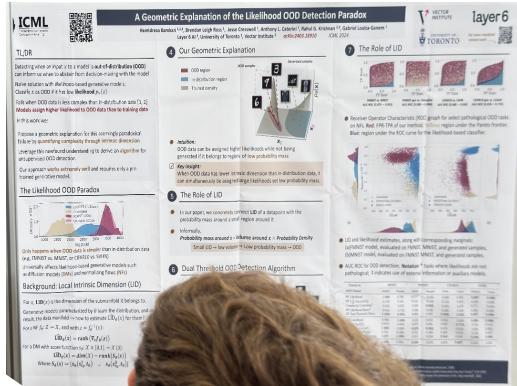
<http://dx.doi.org/10.1103/PhysRevE.91.022103>

Continued on next page

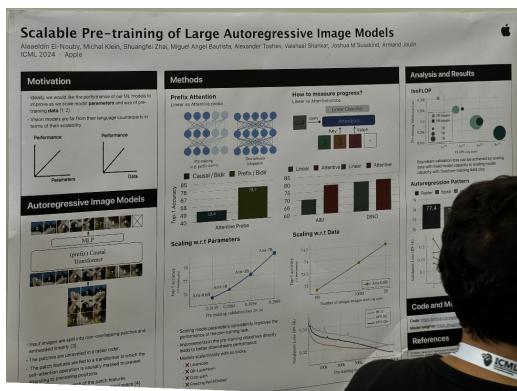
Table 1: (Continued)



Adversarial Perturbations Cannot Reliably Protect Artists From Generative AI *Robert Höning, Javier Rando, Nicholas Carlini, Florian Tramèr*
Adversarial Perturbations Cannot Reliably
<http://arxiv.org/abs/2406.12027v1>



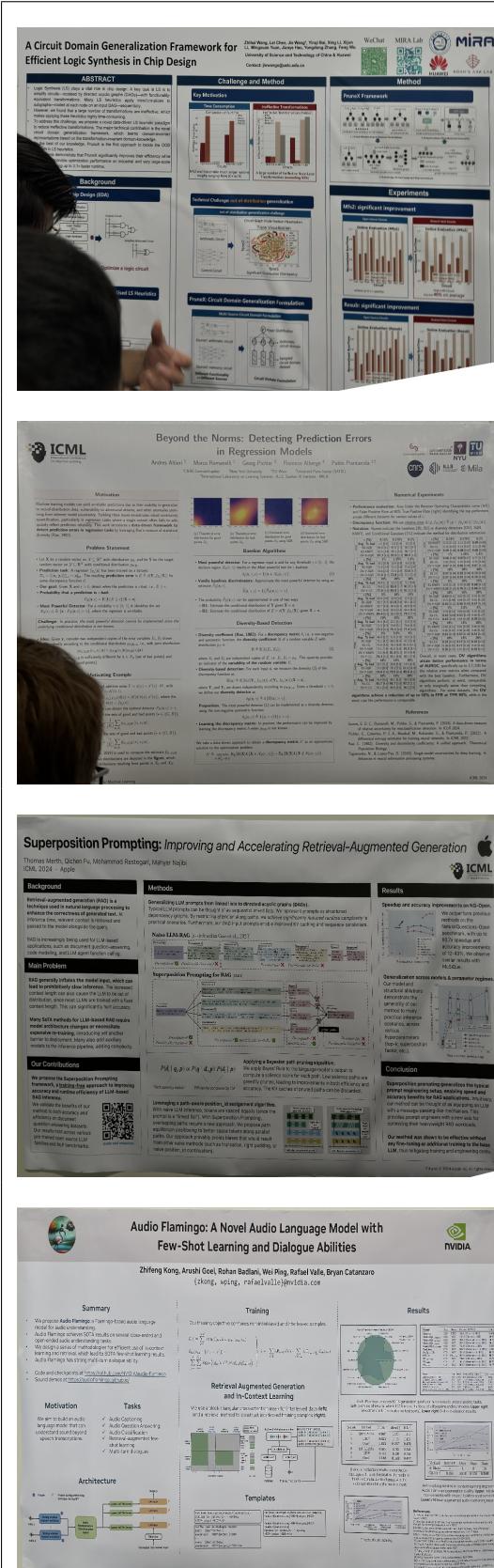
A Geometric Explanation of the Likelihood OOD Detection Paradox *Hamidreza Kamkari, Brendan Leigh Ross, Jesse C. Cresswell, Anthony L. Caterini, Rahul G. Krishnan, Gabriel Loaiza-Ganem*
A Geometric Explanation of the Likelihood OOD Detection Paradox
<http://arxiv.org/abs/2403.18910v2>



Scalable Pre-training of Large Autoregressive Image Models *Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Angel Bautista, Alexander Toshev, Vaishaal Shankar, Joshua M Susskind, Armand Joulin*
Scalable Pre-training of Large Autoregressive Image Models
<http://arxiv.org/abs/2401.08541v1>

Continued on next page

Table 1: (Continued)



A Circuit Domain Generalization Framework for Efficient Logic Synthesis in Chip Design Zhihai Wang, Lei Chen, Jie Wang, Xing Li, Yinqi Bai, Xijun Li, Mingxuan Yuan, Jianye Hao, Yongdong Zhang, Feng Wu
Efficient Logic Synthesis in Chip Design
<http://arxiv.org/abs/2309.03208v1>

Foliations on double-twisted products André Gomes

2122202171071721
<http://arxiv.org/abs/1101.5730v1>

Superpositions of thermalisation states in relativistic quantum field theory Joshua Foo, Magdalena Zych

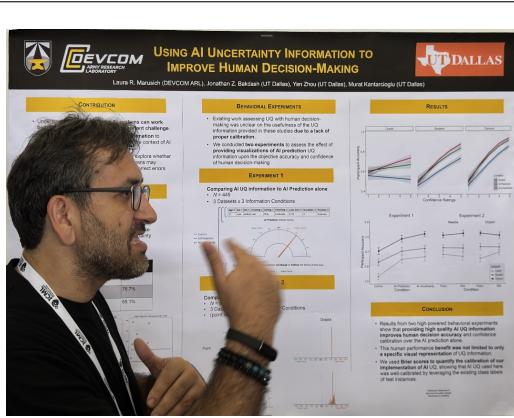
Superposition Prompting: Improving and Accelerating Retrieval-Augmented Generation
<http://arxiv.org/abs/2307.02593v1>

Zero-shot audio captioning with audio-language model guidance and audio context keywords Leonard Salewski, Stefan Fauth, A. Sophia Koepke, Zeynep Akata

Audio Flamingo: A Novel Audio Language Model with
<http://arxiv.org/abs/2311.08396v1>

Continued on next page

Table 1: (Continued)



Using AI Uncertainty Quantification to Improve Human Decision-Making

Laura R. Marusich (DEVCOM ARL), Jonathan Z. Bakdash (UT Dallas), Yan Zhou (UT Dallas), Murat Kantarcioglu (UT Dallas)

Laura R. Marusich (DEVCOM ARL), Jonathan Z. Bakdash (UT Dallas), Yan Zhou (UT Dallas), Murat Kantarcioglu (UT Dallas)

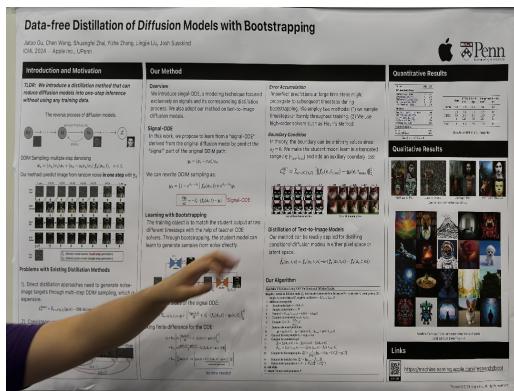
<http://arxiv.org/abs/2309.10852v2>



About Geometry and Initial Phase of Cloud-to-Ground Lightning

Aleš Berkopěc Crafto

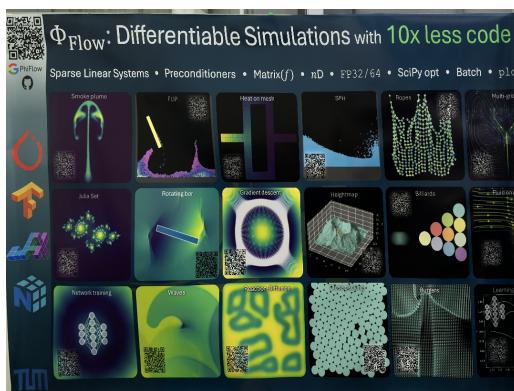
<http://arxiv.org/abs/1602.02496v1>



Data-free Distillation of Diffusion Models with Bootstrapping

fo (xt, t, c) = fox, t, n) + w. (fo (xt, t, c) - foxt, t, n))

Data-free Distillation of Diffusion Models with Bootstrapping



Measuring the Earth's Synchrotron Emission from Radiation Belts with a Lunar Near Side Radio Array

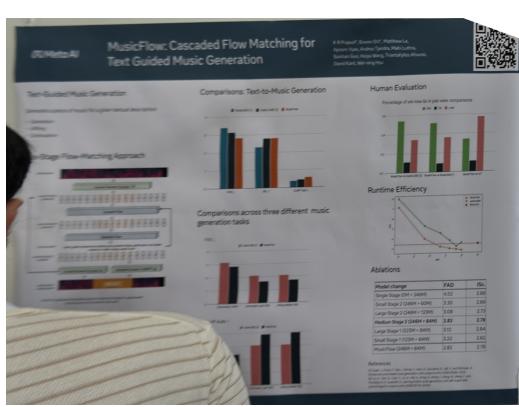
Alexander Hegedus, Quentin Nenon, Antoine Brunet, Justin Kasper, Angelica Sicard, Baptiste Cecconi, Robert MacDowall, Daniel Baker

\$ Flow: Differentiable Simulations with 10x less code

<http://dx.doi.org/10.1029/2019RS006891>

Continued on next page

Table 1: (Continued)



A double-layer Boussinesq-type model for
highly nonlinear and dispersive waves *Florent
Chazel, Michel Benoit, Alexandre Ern, Serge Piperno*
MusicFlow: Cascaded Flow Matching for
<http://dx.doi.org/10.1098/rspa.2008.0508>

WebLINX *WebLINX: Real-World [...]*

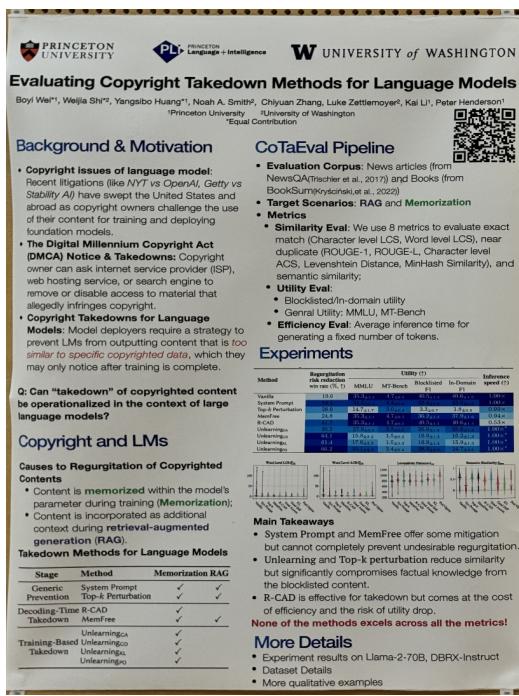


Evaluating Copyright Takedown Methods for Language Models

Boyi Wei, Weijia Shi, Yangsibo Huang, Noah A. Smith, Chiyuan Zhang, Luke Zettlemoyer, Kai Li, Peter Henderson

Evaluating Copyright Takedown Methods for Language Models

<http://arxiv.org/abs/2406.18664v3>



Continued on next page

Table 1: (Continued)

OAK: Enriching Document Representations using Auxiliary Knowledge for Extreme Classification
Shikhar Mohan*, Deepak Sami*, Anirudh Mitra*, Sayali Ray Choudhury*, Bhavana Patel*, Jan Jen*, Manish Gupta*, Manik Varma*
Microsoft Research, India | Microsoft Redmond, US | Microsoft DC, India

Auxiliary Knowledge-Induced Learning for Automatic Multi-Label Medical Document Classification Xindi Wang, Robert E. Mercer, Frank Rudzicz

OAK: Enriching Document Representations using Auxiliary Knowledge for Extreme Classification
<http://arxiv.org/abs/2405.19084v1>

SECOND-ORDER UNCERTAINTY QUANTIFICATION: A DIVERGENCE-BASED APPROACH
Yiyan Li*, Yihui He*, Ming Tang*, and Fei Tian*
TUM Institute of Machine Learning, TUM School of Management

Power-Law distributions and Fisher's information measure F. Pennini, A. Plastino

SECOND-ORDER UNCERTAINTY QUANTIFICATION:
<http://dx.doi.org/10.1016/j.physa.2003.10.076>

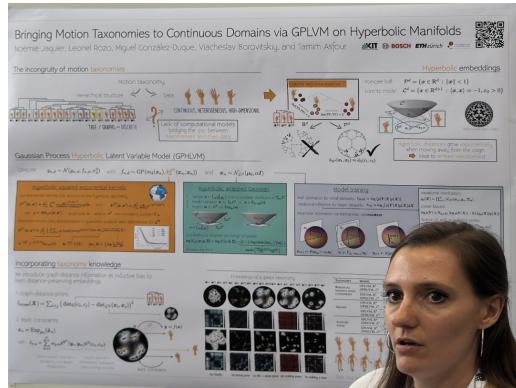
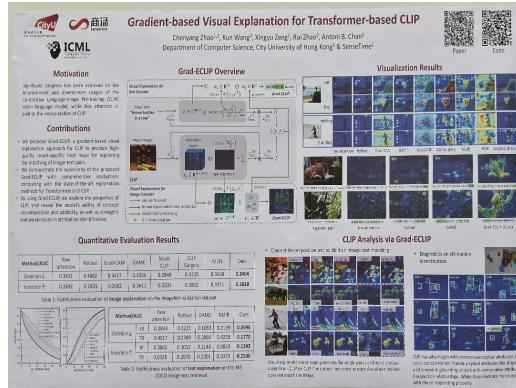
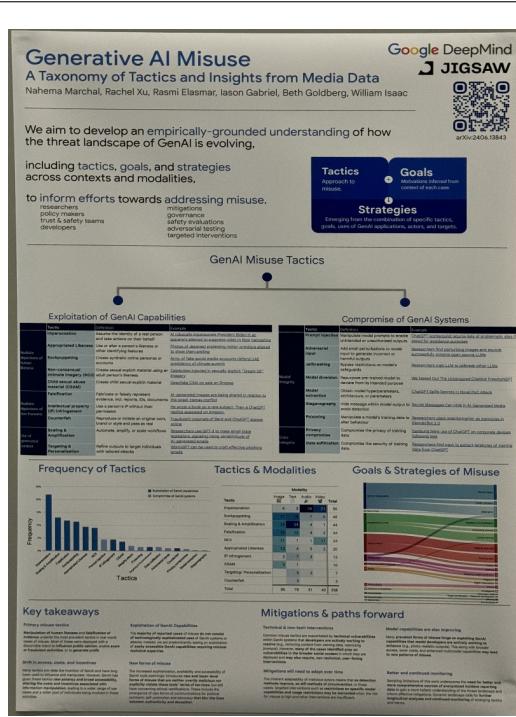
Revisiting the Role of Language Priors in Vision-Language Models
Zhiqiu Lin*, Xinyue Chen*, Deepak Pathak, Penghuang Zheng, Deva Ramanan
ICML 2023, August 2023, Virtual, USA

Revisiting the Role of Language Priors in Vision-Language Models [1] Yuksekgonul et al. (2023). "When and why vision-language models behave like bags-of-words, and what to do about it?", In: ICLR.

Revisiting the Role of Language Priors in Vision-Language Models

Continued on next page

Table 1: (Continued)



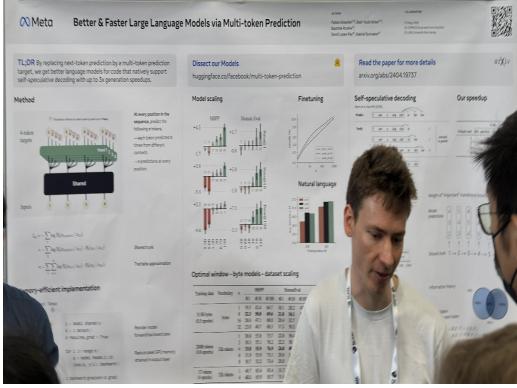
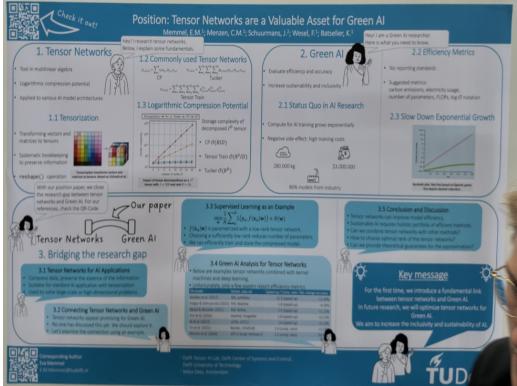
Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data
Nahema Marchal, Rachel Xu, Rashi Elasmar, Jason Gabriel, Beth Goldberg, William Isaac
A Taxonomy of Tactics and Insights from Media Data
<http://arxiv.org/abs/2406.13843v2>

RCA: Region Conditioned Adaptation for Visual Abductive Reasoning *Hao Zhang, Yeo Keat Ee, Basura Fernando*
Gradient-based Visual Explanation for Transformer-based CLIP
<http://arxiv.org/abs/2303.10428v4>

Bringing motion taxonomies to continuous domains via GPLVM on hyperbolic manifolds *Noémie Jaquier, Leonel Rozo, Miguel González-Duque, Viacheslav Borovitskiy, Tamim Asfour*
Bringing Motion Taxonomies to Continuous Domains via GPLVM on Hyperbolic Manifolds
<http://arxiv.org/abs/2210.01672v4>

Continued on next page

Table 1: (Continued)

	<p>Better & Faster Large Language Models via Multi-token Prediction <i>Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Rozière, David Lopez-Paz, Gabriel Synnaeve</i> Better & Faster Large Language Models via Multi-token Prediction http://arxiv.org/abs/2404.19737v1</p>
	<p>TJ-FlyingFish: Design and Implementation of an Aerial-Aquatic Quadrotor with Tilttable Propulsion Units <i>Xuchen Liu, Minghao Dou, Dongyue Huang, Biao Wang, Jinqiang Cui, Qinyuan Ren, Lihua Dou, Zhi Gao, Jie Chen, Ben M. Chen</i> University of Science and Technology of China Microsoft Research & Microsoft Azure The Chinese University of Hong Kong, Shenzhen http://arxiv.org/abs/2301.12344v2</p>
	<p>Position: Tensor Networks are a Valuable Asset for Green AI <i>Eva Memmel, Clara Menzen, Jetze Schuurmans, Frederiek Wesel, Kim Batselier</i> Position: Tensor Networks are a Valuable Asset for Green AI http://arxiv.org/abs/2205.12961v2</p>

Continued on next page

Table 1: (Continued)

INTRODUCTION
Problems Statement: Different tasks require different explanations. However, most existing methods for producing explanations do not have ways to optimize for specific properties.

Contribution: We propose to directly optimize explanations for desirable properties.
 $W_{\text{opt}} = \arg\max_{W} L(W; \text{Prop}_1, \dots, \text{Prop}_k, A)$

where W is a feature-attribution explanation; Prop_i is an explanation Property; A is the set of hyperparameters.

Benefits of our method:

1. Closed-form expression for properties guarantees us for global optimal solution
2. The trade-off of hyperparameter A allows for explicit trade-off when there is a tension between properties

EXPERIMENTAL SETUP
Properties: We consider formulations of two commonly studied properties, faithfulness and robustness:
 $\text{Faith_Loss}(W_F) = \sum_{n=1}^N \|W_{G_n} - W_{R_n}\|_2^2$ (1)
 $\text{Robust_Loss}_k(W_F) = \sum_{n=1}^N \|W_{G_n} - W_{R_n}\|_k \|b_n\|_k$ (2)

where $W_{G_n} = \nabla f(x_n)$ is the "ground-truth" explanation for x_n and $R_n = kx_n + e'_n$ for a general function f .

Property-Optimized Explanations: We generate explanations $W_{\text{opt},k}$ by optimizing the following Loss function:

$$\sum_{n=1}^N \|\nabla W_{G_n} - \nabla W_{R_n} - \nabla W_{\text{opt},k}^k \cdot x_n\|^2 + \lambda \left(\sum_{n=1}^N \sum_{k=1}^K \|W_{G_n} - W_{R_n}\|_k^2 + R(x_n, e'_n) \right)$$

Results: We consider two methods for local feature-based explanations: LIME and SmoothGrad. We match their sampling distribution to the kernel function R chosen in the following way:

EXPERIMENTAL RESULTS
We compare the robustness and faithfulness of SmoothGrad, LIME, and our proposed method, for a range of hyper-parameters. We are looking for how well each explanation generates the following:

1. optimizes for robustness and faithfulness
2. trade-offs between the two properties, while they can both be optimized simultaneously

Below, we see faithfulness and robustness of explanations for functions $f(x) = \sum_{i=1}^d x_i^2$ (left) and $f(x) = \sum_{i=1}^d x_i^2 + \sin(3x_i)$ (right). SmoothGrad and LIME are explained with different levels of faithfulness, but fail to provide a robust explanation. Our proposed method can achieve both robustness and faithfulness simultaneously. On the right, we see that hyper-parameters of LIME and SmoothGrad produce no variations in either faithfulness or robustness along λ . On the left, we see that our explanation method offers trade-offs.

DISCUSSION
We propose a directly optimizing local feature-based explanations, it is able to match properties. By doing so, our method can explicitly manage trade-offs between properties through hyperparameters. This is in contrast to the process of directly optimizing explanation for desirable properties because: (1) prior works show that it is hard to find a good explanation for different downstream tasks [5], and (2) optimized explanations can provide a better starting point for user studies on XAI [8].

ONGOING WORK
We are building on this work by:

- Including more properties as well as different formulations of faithfulness and robustness
- Additional baseline explanation generation methods
- Drawing the impact of our property-optimal explanations on real-world datasets and tasks
- Scaling our optimization method.

ACKNOWLEDGEMENTS
This material is based upon work supported by the National Science Foundation under Grant No. IIS-1760038. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

- [1] Zhou, Jianlong, Gandomi, Amir H., Chen, Feng, and Behzadian, Mohammad. Optimizing the quality of machine learning explanations: A survey on explainability in machine learning. *EURASIP J. Adv. Signal Process.*, 2019.
- [2] Zhou, Jianlong, Gandomi, Amir H., Chen, Feng, and Behzadian, Mohammad. Explainability in machine learning: A review. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2018.
- [3] Zhou, Jianlong, Gandomi, Amir H., Chen, Feng, and Behzadian, Mohammad. Explainability in machine learning: A review. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2018.
- [4] Zhou, Jianlong, Gandomi, Amir H., Chen, Feng, and Behzadian, Mohammad. Explainability in machine learning: A review. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2018.
- [5] Zhou, Jianlong, Gandomi, Amir H., Chen, Feng, and Behzadian, Mohammad. Explainability in machine learning: A review. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2018.
- [6] Zhou, Jianlong, Gandomi, Amir H., Chen, Feng, and Behzadian, Mohammad. Explainability in machine learning: A review. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2018.
- [7] Zhou, Jianlong, Gandomi, Amir H., Chen, Feng, and Behzadian, Mohammad. Explainability in machine learning: A review. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2018.
- [8] Zhou, Jianlong, Gandomi, Amir H., Chen, Feng, and Behzadian, Mohammad. Explainability in machine learning: A review. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2018.

EXPAND AND CLUSTER
Parameter Recovery of Neural Networks
Flavio Martinelli*, Berlin Šimşek*, Wulfraam Gersmeyer*, Johanni Breu*

1. Can we identify NN weights only from input-output queries?

2. Reaching zero training loss is a hard problem. We apply over-parametrization to reduce non-convexity.

3. Through symmetry, teacher weights emerge. We match them between student and teacher.

4. Expand-and-Cluster recovers weight and layer sizes algorithm and results on feed-forward nets

EPFL
Flavio.Martinelli@epfl.ch
Flavio.Martinelli.github.io

1) Train STUDENTS to imitate logic of a TRACER-ER network
2) NOT CONVEX loss landscape? OVER-PARAMETERIZED!
3) Know their SYMMETRIES? Weights of any two hyperparameters must be identical in structure
4) CLUSTER student weight vectors to expose the teacher's

Dual Convexified Convolutional Neural Networks
Site Bai, Chuyang Ke, Jean Honorio
Parameter Recovery of Neural Networks
<http://arxiv.org/abs/2205.14056v2>

Creative Text-to-Audio Generation via Synthesizer Programming
Michael Charley, Harish Singh, Jessica Bland

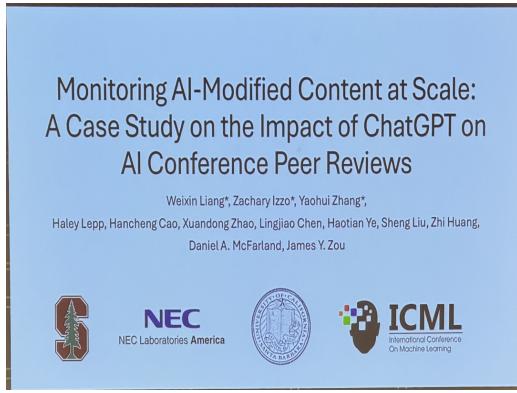
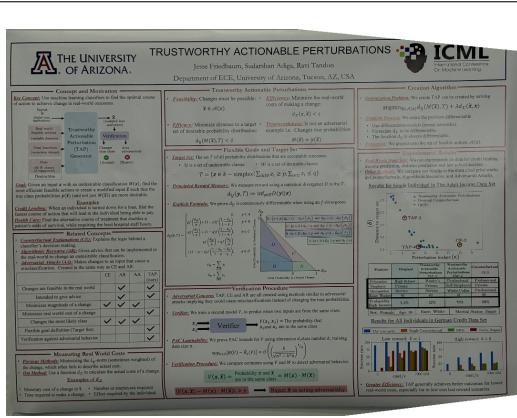
ICML 2024, New York City, NY, USA
July 14–19, 2024
Poster presentation

Abstract: Generating creative audio from text is a challenging task. We introduce a novel approach for generating creative audio by leveraging the expressive power of synthesizers. Our method involves generating a sequence of notes and velocities for a given text input, which are then processed by a synthesizer to produce the final audio output. We demonstrate the effectiveness of our approach by generating creative audio from various text inputs, including poems, stories, and songs. Our results show that our approach can generate audio that is both musically interesting and contextually appropriate.

Luban: Building Open-Ended Creative Agents via Autonomous Embodied Verification
Yuxuan Guo, Shaohui Peng, Jiaming Guo, Di Huang, Xishan Zhang, Rui Zhang, Yifan Hao, Ling Li, Zikang Tian, Mingju Gao, Yutai Li, Yiming Gan, Shuai Liang, Zihao Zhang, Zidong Du, Qi Guo, Xing Hu, Yunji Chen
Creative Text-to-Audio Generation Via Synthesizer Programming
<http://arxiv.org/abs/2405.15414v1>

Continued on next page

Table 1: (Continued)



A Survey on Trustworthy Edge Intelligence: From Security and Reliability To Transparency and Sustainability Xiaojie Wang, Beibei Wang, Yu Wu, Zhaolong Ning, Song Guo, Fei Richard Yu

TRUSTWORTHY ACTIONABLE PERTURBATIONS

<http://arxiv.org/abs/2310.17944v2>

Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews

Weixin Liang*, Zachary Izzo*, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xudong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, James Y. Zou

Monitoring AI-Modified Content at Scale:
<http://arxiv.org/abs/2403.07183v2>