# The Israel Statistics and Data Science Association

## 2026 Annual Conference – Tel Aviv

# ABSTRACTS

---

## Session 1-A – Measuring and Steering Large Language Models

**Yonathan Efroni, Tel Aviv University: Hack-Verifiable Environments: Towards Evaluating Reward Hacking at Scale**

As AI systems become increasingly autonomous and capable, ensuring that their behavior aligns with human intent remains a fundamental challenge. One important failure mode is *reward hacking*, where agents exploit shortcuts that maximize evaluation signals while violating the true objective of the task. Although reward hacking has been observed across a wide range of domains, existing evaluation methods remain limited: they are often restricted to narrow settings and typically rely on costly post hoc inspection by humans or LLM judges. This makes it difficult to systematically measure and study reward hacking across diverse environments and agentic settings.

In this talk, we will explore a new idea for evaluating reward hacking that enables reliable and scalable measurement of it. Instead of identifying reward hacking after the execution, we embed verifiable reward hacking opportunities directly into environments, making exploitation detectable by design and enabling deterministic automated evaluation. We will formalize this idea via the framework of Hack-Verifiable Environments and present Hack-Verifiable TextArena, a benchmark suite built on TextArena for measuring reward hacking across diverse single-agent and multi-agent environments. Using this benchmark, we analyze reward hacking behavior across frontier and open-source language models and discuss the broader implications for AI evaluation and alignment.

**Gabriel Stanovsky, Hebrew University: On the Brittleness of LLM Evaluation**

Meaningful evaluation of LLMs is central to scientific progress: it helps us understand model behavior, identify current limitations, and chart directions for improvement. In this talk, I will argue that despite its importance, current evaluation is often brittle, producing inconsistent and even contradictory conclusions. I begin with our recent large-scale statistical analyses, which show that even minimal prompt paraphrases can cause substantial shifts in both absolute performance and the relative ranking of models. I then outline desiderata for robust LLM evaluation, present a set of metrics tailored to different use cases, and conclude by proposing a probabilistic benchmarking framework.

**Rotem Dror, University of Haifa: Deep Persona: A Psychologically Grounded Architecture and Evaluation Framework for Role-Playing Agents and Simulations**

Large Language Models (LLMs) are increasingly deployed for persona simulation, however current approaches rely on persona prompts defined by superficial trait descriptions and stylistic constraints that fail to sustain coherent behavior over extended interactions. We introduce Deep Persona, a psychologically grounded, three-layered architecture for constructing highly convincing role-playing agents. Our methodology is governed by principles of scripted determinism and bounded agency, restricting the LLM to a reactive engine guided by a strict internal script, rather than relying on its unconstrained generative capabilities. To assess persona realism, we propose an evaluation framework comprising reference-free psychological metrics that statistically benchmark dialogue naturalness against human distributions, drawing on established psychological clinical instruments and adversarial stress-tests. Empirical evaluation comparing multiple human-LLM interactions to human-human baselines

reveals that while LLMs achieve high pragmatic fluency, they exhibit systematic limitations in deeper conversational dimensions, such as emotional expression and joint attention.

**Idan Szpektor, Google Research: Exploring AI Memorization**

Just like humans, AI models require various memory mechanisms to help them accomplish different types of tasks, such as thinking about complex requests, memorizing world knowledge and assimilating personal experiences along large time frames. In this talk I will bring up the AI equivalent of the four memory mechanisms in the brain: working memory, procedural memory, semantic memory and episodic memory. Specifically, I will present several works that address aspects of these mechanisms, including ethics in decision making, uncertainty expression in fact recall and episodic memory fusion for long-horizon embodied tasks.

# Session 1-B – Inference in Complex Models

**Yuval Benjamini, Hebrew University: Uncertainty Intervals for Ranking, with Applications to Machine Learning**

In many scientific and machine learning evaluations, stakeholders care more about who ranks first, second, or third than about the exact numerical scores. However, the uncertainty in these rankings, which can be substantial, is often ignored. This talk develops methods for constructing rank intervals with explicit error guarantees using pairwise hypothesis tests. For a single ranking task, rank confidence intervals can be derived from pairwise tests adjusted to control the family-wise error rate. However, these intervals are often overly conservative. We instead use false discovery rate control and analyze the statistical properties and efficiency gains of the resulting intervals. For settings with multiple ranking tasks over the same competitors, such as model leaderboards, we propose an aggregation framework based on conformal prediction intervals. We apply these methods to analyze ranking uncertainty in a public question-answering leaderboard for pre-trained large language models.

The talk is based on joint work with Bitya Neuhof and Yoav Benjamini

**Marina Bogomolov, Technion: Leveraging Hypotheses' Group Structure for Powerful Testing with Post-Filtering FDR Control**

Modern biological studies often involve testing many hypotheses organized in a group or a hierarchical structure, such as a directed acyclic graph (DAG). In such studies, researchers often aim to control the false discovery rate (FDR) after filtering discoveries to obtain interpretable results. Katsevich, Sabatti, and Bogomolov (2023) introduced Focused BH, a procedure that controls the FDR of the filtered rejection set under certain assumptions. We propose improving its power by incorporating data-dependent weights that adapt to group or hierarchical structures, yielding Weighted Focused BH (WFBH). For DAG-structured hypotheses, we propose a variant of WFBH that gains power by adapting to the DAG structure and exploiting logical relationships among hypotheses. We establish post-filtering FDR control for WFBH and its DAG variant under certain assumptions. Simulations show that the DAG variant of WFBH is robust to deviations from these assumptions and can be considerably more powerful than comparable methods. Finally, we demonstrate its practical utility using microbiome and gene expression datasets.
This is joint work with Shinjini Nandi.

**Ruth Heller, Tel Aviv University: Selecting informative conformal prediction sets optimally with an FCR guarantee**

Conformal methods provide prediction sets for outcomes with confidence guarantees. We study their use in a selective inference setting, where inference is performed only when the prediction set is informative. The analyst may consider as informative, for example, cases with prediction sets that are sufficiently small, exclude null values, or satisfy other appropriate monotone constraints. Because inference is typically

restricted to informative cases in practical applications, accounting for the resulting selection bias is crucial to maintaining valid error control. We derive the optimal decision policy under a suitable power objective in the oracle setting where the probability of belonging to each prediction set can be computed. In practice, of course, only estimated probabilities are available. We therefore introduce a calibration procedure that adjusts the oracle policy to maintain finite sample FCR control. We demonstrate the effectiveness of our new methods for classification outcomes on both real and simulated data.

Joint work with Israela Solomon, Etienne Roquain, and Saharon Rosset

**Meshi Bashari, Technion: General Synthetic-Powered Inference**

How can we obtain trustworthy inference while leveraging untrusted synthetic data? In this talk, I will introduce a framework that safely improves the sample efficiency of a broad class of statistical inference procedures---including conformal prediction and hypothesis testing---by adaptively leveraging untrusted synthetic data, such as data generated by modern generative models. Crucially, this framework provides distribution-free error-control guarantees without making any assumptions about the quality of the synthetic data. I will demonstrate its broad applicability across diverse domains, ranging from reliable protein structure prediction to principled win-rate evaluation of large reasoning models. If time permits, I will also discuss how these ideas can be used to improve the sample efficiency of the Benjamini–Hochberg procedure for false discovery rate control without sacrificing its guarantees.

# Session 2-A – Theoretical Frontiers for Modern AI

**Gilad Yehudai, New York University: When Can Transformers Count to n?**

Large language models can solve highly complex tasks, but why do they still struggle to solve simple problems such as counting? In this talk, we show that this isn't just a quirk of training data, it's a hard mathematical limit inherent to the transformer architecture. We reveal a sharp theoretical phase transition governed by the relationship between a model's embedding dimension and its vocabulary size. When the dimension is at least as large as the vocabulary, transformers can perfectly maintain token counts. But the moment the vocabulary outgrows the dimension, tokens become geometrically crowded. We prove that separating these crowded tokens forces the network's weights to scale polynomially, making exact counting numerically unstable and difficult to learn. We will show empirical results on transformers trained from scratch and pretrained LLMs that confirm this exact failure threshold. Ultimately, this exposes a critical blind spot in how we evaluate long-context models, showing that vocabulary size, and not just context length, is a major structural bottleneck for reliable counting.

**Alon Peled-Cohen, Tel Aviv University: Real price of bandit information in multiclass classification**

Does limited feedback fundamentally harden multiclass classification? In bandit multiclass classification, a learner receives data points sequentially and must predict their labels, receiving feedback only on whether their guess was correct. This limited feedback is traditionally thought to increase sample complexity—a cost known as the "price of bandit information." While classic literature suggests this cost scales with the number of labels, it remains unknown if this bound is truly necessary.
In the talk, I will discuss recent results that challenge this assumption, proving that the price of bandit information vanishes asymptotically. We extend this result across both PAC learning and online learning settings, showing that we can achieve the same efficiency with limited feedback as we can with full-information labels.

**Gal Vardi, Weizmann Institute: Theoretical Aspects of Learning with Chain-of-Thought**

To solve complex tasks, language models produce a chain-of-thought leading to the desired answer, where each intermediate token is generated in an autoregressive manner. In this talk, I will present a formal learning framework for studying this emerging paradigm, both when chain-of-thought is observed during training and when only prompt–answer pairs are used, with the chain-of-thought remaining latent. I will discuss the benefits of learning with chain-of-thought examples in terms of sample and computational complexity. Finally, I will discuss learning with chain-of-thought supervision from multiple thinkers, each

providing correct but possibly systematically different solutions, e.g., step-by-step solutions to math problems written by different thinkers.

**Shay Moran, Technion: Three Frameworks for Generalization in Learning Theory**

We will discuss several related mathematical frameworks for generalization in supervised learning, with a focus on classification. We begin with the classical PAC framework of Vapnik–Chervonenkis and Valiant, which provides a clean and powerful worst-case theory of learning. While highly successful, this framework does not fully capture some aspects of modern practice, where algorithms often generalize far better than worst-case guarantees suggest.

I will then present two seemingly modest refinements of the PAC model that go a long way. These allow us to capture, in a clean mathematical way, situations in which the data distribution has "nice" structure that facilitates learning, and help explain when and why faster learning rates are possible.

In contrast to classical PAC learning – whose algorithmic landscape is by now well understood – these refined frameworks remain largely unexplored. I will conclude, time permitting, with open questions highlighting what we do not understand about algorithms and optimal learning rates in these settings.

# Session 2-B – Statistical Modeling

**Ori Davidov, Haifa University: Least squares for cardinal paired comparisons data**

Least square estimators for graphical models for cardinal paired comparison data with and without covariates are rigorously analyzed. Novel, graph--based, necessary and sufficient conditions that guarantee strong consistency, asymptotic normality and the exponential convergence of the estimated ranks are emphasized. A complete theory for models with covariates is laid out. In particular, conditions under which covariates can be safely omitted from the model are provided. The methodology is employed in the analysis of both finite and infinite sets of ranked items where the case of large sparse comparison graphs is addressed. Lack of fit tests are developed. The proposed methods are explored by simulation and applied to the ranking of teams in the National Basketball Association (NBA).

**Amit Donner, Haifa University: Statistical inference in circular structural model and fitting circles to noisy data**

It is well known that commonly used algorithms for circle fitting perform poorly when sampling distribution of the points is not symmetric with respect to the circle center, e.g., when the points are sampled from a circle arc. To overcome this difficulty we introduce and study a parametric circular structural model. In this model the points are assumed to be sampled according to the von Mises distribution with unknown concentration and mean direction parameters. Under these circumstances we develop maximum likelihood and method of moments estimators of the circle center and radius, and study their statistical properties. In particular, we show that the proposed maximum likelihood estimator is asymptotically normal and efficient. We also develop a test of uniformity for the sampling distribution along the circle. Based on the derived theoretical results we propose a numerically stable circle fitting algorithm, investigate its accuracy in a simulation study, and illustrate its behavior in a real data example.

**Malka Gorfine, Tel Aviv University: Flexible Deep Neural Networks for Partially Linear Survival Data: Estimation and Survival Inference**

We propose a flexible deep neural network (DNN) framework for modeling survival data within a partially linear regression structure. The approach preserves interpretability through a parametric linear component for covariates of primary interest, while a nonparametric DNN component captures complex time–covariate interactions among nuisance variables. We refer to the method as FLEXI–Haz, a FLEXIble Hazard model

with a partially linear structure. In contrast to existing DNN approaches for partially linear Cox models, FLEXI–Haz does not rely on the proportional hazards assumption. We establish theoretical guarantees: the neural network component attains minimax-optimal convergence rates over composite Hölder classes, the linear estimator is sqrt-n-consistent, asymptotically normal, and semiparametrically efficient, and we develop a cross-fitted one-step estimator of the cumulative hazard and survival function for a new subject, together with pointwise asymptotic confidence intervals. To the best of our knowledge, this is the first frequentist asymptotic pointwise inference result for a survival function in a DNN survival model, with or without a linear component. Simulations and real-data analyses demonstrate the utility of FLEXI–Haz as a principled and interpretable alternative to methods based on proportional hazards. Code for implementing FLEXI-Haz, as well as scripts for reproducing data analyses and simulations is available on GitHub site https://github.com/AsafBanana/FLEXI-Haz. This is a joint work with Asaf Ben Arie.

**Tamir Zehavi, Tel Aviv University: Regression discontinuity designs with truncation by death**

Regression Discontinuity (RD) methodologies have received increasing attention in recent years. Although these methods are widely used for causal inference, several complications remain understudied. In this work, we address one such complication by extending the RD framework to accommodate a fundamental challenge arising across various contexts: the truncation of outcomes due to death. We employ principal stratification to target the survivor average causal effect within compliers (c-SACE) and provide an identification strategy and a nonparametric estimation procedure for this setting. While previous studies have primarily focused on developing bounds for the c-SACE, we derive point identification of this estimand, in both sharp and fuzzy RD designs, by adopting a covariate-conditional version of the RD assumptions. Importantly, conditioning on a rich set of covariates enhances the plausibility of the RD assumptions.

Moreover, in contrast to much of the existing literature, which typically relies on continuity of the covariate distribution around the RD threshold, our approach allows identification of RD causal effects even when the covariate density is discontinuous at the threshold. We conduct simulation studies and apply our method to a real dataset. Because key identifying assumptions are cross-world and cannot be tested empirically, we develop sensitivity analysis techniques and demonstrate their utility in our application.