TPC® Newsletter

Issue 6, December 2025

TPC Technology Conference 2025
TPCTC London, United Kingdom, September 1, 2025

18 Research Submissions Power TPCTC25's Most Diverse Program Yet.

TPCTC is TPC's annual technology conference. Its mission is to bring together industry experts and researchers to explore new methodologies for measuring the performance of data-centric applications. Over the last 16 years TPCTC has been recognized as the international event for anyone interested in performance related topics in database technology, including Transaction Processing, Data Warehousing, Big Data Analytics, Internet of Things, Virtualization, Hyper-Converged Infrastructure, and Artificial Intelligence.

Benchmark Innovation

Performance evaluation has been and still is one of the main differentiators of computer systems. Constant hardware and software improvements require frequent evaluation of existing performance methodologies to allow for them to be technically sound, fair, and meaningful. The TPC has always been at the forefront of these developments. While the TPC has focused historically on database centric benchmarks, recent developments include benchmarks for Artificial Intelligence, Internet-Of-Things, Hyper-Converged Infrastructure, Big Data and Virtualization. Many of these benchmarks were sparked by ideas that originated in papers presented at past TPCTC events. These papers were mostly academic papers that inspired ideas for new benchmarks, identified deficiencies in existing benchmarks, and motivated improvements.

At the same time, the academic community and industry have benefited from TPCTC as they define their own performance methodologies. The last ten years have seen the rise of many new DBMSs, some with unique approaches to traditional problems, like columnar databases, in-memory databases, others specializing on specific applications, such as graph databases, NoSQL databases, Timeseries databases and document databases etc. With them came a flurry of new performance methodologies resulting in customized benchmarks, many of which were based on methodologies originally developed in the TPC.

TPCTC has always served as a venue where both practitioners with real world performance expertise and highly innovative academics meet to discuss performance methodologies for emerging technologies. The results, both in academia and TPC benchmarks, are a direct result of this knowledge exchange.

TPCTC25, chaired by Raghu Nambiar and Meikel Poess, took place in London, United Kingdom, September 1, 2025, in conjunction with the 51st conference on Very Large Databases (VLDB), a top tier academic conference for databases. TPCTC25 marks TPCTC's 17th anniversary.

Starting last year TPCTC has been offering authors the option to submit the tools and scripts used to conduct performance measurements together with their papers. The TPC refers to the tools and scripts as a workload artifact. Accepted workload artifacts will be hosted in TPC's GitHub repository and promoted on the TPC's website together with author names. The TPC will also provide a mechanism for results from workload artifacts to be listed on the TPC's webpage.

This year we received 18 regular paper submissions in a variety of research topics and accepted 10 papers in the following research areas:

Core Database Benchmarking (4 Papers):

This is the largest group, focusing on creating or refining benchmarks for different types of data systems. Topics include new benchmarks for databases with varying value lengths, JSON data analytics (CH2++), and transactional key-value stores (Tectonic). A specific paper proposes benchmarking Role-Based Access Control.

AI/LLM-Related Benchmarking and Performance (3 Papers):

This group addresses the emerging need to benchmark AI workloads. Submissions cover flexible LLM inference benchmarking across different hardware (ScaleBench_AI), evaluating distilled language models, and reporting on the journey to achieving leadership performance in MLPerf submissions.

Data Generation and Augmentation (2 Papers):

These papers focus on creating realistic test data for benchmarking. One proposes a system for generating synthetic relational data from SQL schemas (DataGenX), and another looks at tabular data augmentation, specifically in the context of medical insurance claims for scalability testing.

System/Hardware Optimization and Sustainability (1 Paper):

This group includes papers that address system-level performance considerations outside of the core data structure. A submission on DiStash focuses on a disaggregated, multi-stash architecture for a transactional key-value store.

The following are summaries of the accepted papers:

Core Database Benchmarking (4 Papers):

A Benchmark for Databases with Varying Value Lengths (First author Danushka Liyanage)

Danushka Liyanage is a Postdoctoral Research Associate in the School of Computer Science at the University of Sydney, Australia. His research interests are centered on quantitative decision-making for software testing techniques and the development of rigorous benchmarking methodologies for modern data management systems, such as the one described in the paper.

The paper introduces a novel benchmarking approach called YCSB-IVS (YCSB Increasing Value Sizes) to address a gap in traditional database benchmarking: the lack of evaluation for workloads where record values grow significantly in length over time. The authors extend the popular Yahoo! Cloud Serving Benchmark (YCSB) by adding an "extend" operation that simulates the appending of data to record fields, causing them to grow. The benchmark methodology involves a sequence of epochs, where an extend phase is followed by a run phase (standard YCSB query workload), allowing the measurement of performance degradation as value sizes increase. By comparing results from the main-run (with history of growth) against a clean-run (same logical state but no history) and baselines (uniform record sizes), the benchmark is designed to disentangle the performance impact of overall record length, the variation in record length, and the historic growth in length. The application of YCSB-IVS to three popular DBMS backends—MongoDB, MariaDB+InnoDB, and MariaDB+MyRocks—revealed significant performance differences tied to the storage engine design and its handling of variable-sized values, underscoring the need for more representative benchmarks.

CH2++: New HOAP for Benchmarking JSON Data Analytics (First Author M.J. Carey)

M.J. Carey is affiliated with Couchbase, Inc., based in San Jose, CA, USA. This work focuses on benchmarking JSON document databases, aligning with Couchbase's specialization in the NoSQL and hybrid operational/analytical processing (HOAP) markets.

The paper introduces CH2++, a major revision of the existing CH benchmark, designed specifically for evaluating Hybrid Operational/Analytical Processing (HOAP) systems that use JSON analytics. This new benchmark addresses the need for a collaborative standard in the document database space, similar to how the relational CH-benCHmark serves the hybrid operational/transaction processing HTAP space.

CH2++ significantly revamps the original schema, adopting a more document-oriented, JSON structure that reflects modern database design best practices. This new schema models structures using nested objects for addresses and names, and nested arrays for customer addresses, phones, and item categories. Critically, it includes 64 "extra fields" in key tables to simulate real-world data width and test a system's ability to efficiently ignore irrelevant columns (a feature vital for columnar storage).

The paper's analytical queries are re-expressed in SQL++ to handle these complex JSON structures. Results running CH2++ on Couchbase's Capella services demonstrated:

- Effective Performance Isolation: The transactional workload remained robust despite concurrent analytical queries.
- Linear Query Speedup: Doubling the analytical cluster size resulted in approximately a halving of the geometric mean query execution time.
- Storage Format Value: The benchmark was instrumental in showing that using a column-oriented JSON storage format resulted in a factor-of-two improvement in query performance over row-oriented format, due to efficient I/O avoidance for unreferenced columns.

Tectonic: Bridging Synthetic and Real-World Workloads for Key-Value Benchmarking (First author Alexander H. Ott)

Alexander H. Ott is a researcher at Brandeis University, Massachusetts, USA. This work stems from an academic context focused on key-value systems research, aiming to bridge the gap between idealized synthetic benchmarks and complex real-world production workloads.

The paper introduces Tectonic, a Rust-based, highly configurable, and resource-efficient key-value (KV) workload generator designed to overcome limitations in existing benchmarks like YCSB and KVBench.

Tectonic models three critical aspects of production workloads ignored by predecessors:

- Dynamic Workloads: Tectonic can model workloads where the operation mix, distribution, and access patterns change arbitrarily over time (e.g., modeling diurnal shifts or seasonal trends in key access).
- Structured Key Generation: It supports customizable composite key generation and prefix-based access patterns (e.g., encoding metadata like relation names into keys), which is crucial for systems that rely on prefix scans for efficiency.
- Variable Data Sortedness: Unlike prior tools, Tectonic can generate data with user-specified degrees of near-sortedness, a characteristic that significantly impacts ingestion, indexing, and query costs in real-world scenarios.

Tectonic achieves these capabilities while being significantly more efficient:

- Speed and Throughput: It generates workloads up to 10-12x faster and achieves 2x higher throughput than state-of-the-art generators.
- Resource Efficiency: It records an average of 7.5x lower main memory footprint (up to 84% less) than state-of-the-art generators, enabling the generation of larger workloads.

Benchmarking Role-Based Access Control in Data Management Systems (First author: Mads Cornelius Hansen)

Mads Cornelius Hansen is a student at the IT University of Copenhagen, Denmark. This paper represents academic research focusing on a key area of database systems architecture and security performance.

The paper proposes the first standardized benchmark for evaluating the performance of Role-Based Access Control (RBAC) mechanisms in data management systems. RBAC is critical for scalable security, but its internal implementation varies widely, leading to inconsistent performance at scale.

The benchmark centers on the role hierarchy—a core feature that can grow to over 100,000 roles in industry applications. The benchmark is split into two components:

1. Role Hierarchy Creation: Measures the performance of iterative CREATE ROLE and GRANT statements when building deep linear, wide linear, and balanced tree structures.
2. Role Hierarchy Access: Measures the performance of accessing data and metadata using standard SQL commands (SELECT * FROM table and SHOW ROLES) to assess authorization overhead and metadata retrieval efficiency.

The evaluation across PostgreSQL, MariaDB (OLTP), and Snowflake (OLAP) revealed significant architectural dependencies:

- PostgreSQL suffered massive performance degradation in the cloud compared to on-prem, attributed to high I/O demands of the GRANT statement conflicting with cloud IOPS limits.
- MariaDB experienced performance degradation during GRANT operations depending on the hierarchy's shape (balanced was best), explained by how it recursively invalidates and rebuilds its privilege cache.
- Snowflake exhibited uniformly high latency for metadata operations due to the overhead of network coordination in its distributed service layer where metadata is stored.

The study concludes that these non-uniform behaviors highlight the urgent need for a standardized RBAC benchmark to facilitate future optimization efforts.

AI/LLM-Related Benchmarking and Performance (3 Papers):

ScaleBench_AI - Flexible LLM Inference Benchmarking Across Architectures and Environments (First author Karthik Krishna)

Karthik Krishna is affiliated with Infobell IT Solutions Pvt. Limited in Bengaluru, India. This affiliation ties the work to the field of AI engineering and infrastructure, focusing on creating tools for optimizing and scaling LLM deployments for enterprises.

The paper introduces ScaleBench_AI, a flexible, dynamic benchmarking framework for evaluating Large Language Model (LLM) inference performance under realistic, production-ready loads. It positions itself as an alternative to hardware-vendor benchmarks like MLPerf, offering capabilities essential for practitioners deploying LLMs across diverse architectures.

ScaleBench_AI is platform- and model-agnostic, capable of testing any LLM model with any inference server (e.g., vLLM, Triton, SGLang) on single-node or multi-node systems. It measures performance using industry-aligned metrics and stringent validation criteria.

The key technical contribution is the automated load threshold discovery mechanism. This system:

- Conducts iterative load testing by gradually increasing the number of concurrent users.
- Uses binary search backtracking when a threshold violation is detected to efficiently pinpoint the optimal user count (maximum concurrency) that still maintains compliant latency.
- Qualifies the final result with 10 test runs to ensure the capacity is statistically reliable.

The framework uses Dataset-20k, a curated and token-bucketed dataset, to simulate varied and realistic prompt lengths. By only counting throughput when validation thresholds are met, ScaleBench_AI provides actionable capacity planning insights for production environments, enabling informed decisions on scaling and provisioning.

Benchmarking Distilled Language Models: Performance and Efficiency in Resource-Constrained Settings (First author Sachin Gopal Wani)

Sachin Gopal Wani is affiliated with Lenovo, Infrastructure Solutions Group in Morrisville, NC, USA.

The paper benchmarks the efficiency and performance of distilled language models to validate knowledge distillation as a strategic approach for creating powerful, efficient Small Language Models (SLMs) for resource-constrained environments.

The study uses total FLOPs (Floating-Point Operations) as the primary metric, converting this cost into practical GPU-Hours and CO_2 emissions. The key finding is that distillation doesn't just save money; it creates a new, superior performance-to-compute curve.

Comparing an 8B vanilla model (trained from scratch) with an 8B distilled model, the results showed:

- Compute Cost Reduction: Distillation required over 2,000 times less compute (or 99.9% less FLOPs) than full pre-training (622 H100 GPU-hours vs. 1.26 million H100 GPU-hours).
- Performance Gain: The distilled model achieved reasoning scores on benchmarks like AIME 2024 and GPQA that were on par with, or exceeded, models (e.g., Llama 3.3 70B) requiring orders of magnitude more compute. For instance, a distilled 8B model achieved higher AIME accuracy than a 235B vanilla model that cost nearly 60,000 times more to develop.

The paper concludes that distillation is a powerful democratizing force in AI, offering a viable, sustainable, and highly compute-efficient pathway to achieve competitive, state-of-the-art reasoning capabilities without demanding elite hardware resources.

Delivering MLPerf Submissions: Journey to Leadership Performance (First author: Miro Hodak)

Miro Hodak is part of the AMD Artificial Intelligence Group located in Austin, Texas, USA.

The paper summarizes AMD's successful year-long effort to publish competitive, and in some cases, leadership performance results on the industry-standard MLPerf AI benchmark, covering both Inference and Training on several AMD Instinct GPU platforms. This achievement was made despite the high barrier to entry for new organizations in MLPerf. The paper outlines the journey, which involved focusing on key sub-benchmarks like the Llama2-70b model and utilizing extensive optimizations, including FP8 quantization, GEMM tuning, and Flash Attention. Beyond the published scores, the authors emphasize that MLPerf provides significant value by improving relationships with server OEMs and cloud providers, generating publicly available and reproducible performance data, supplying data for RFPs and customer requests, and providing technical marketing opportunities. They conclude that success in MLPerf requires deep optimization, cross-company alignment, and continuous improvement to keep pace with the fast-moving field.

Data Generation and Augmentation (2 Papers):

DataGenX: Generating Synthetic Relational Data from Annotated SQL Schemas (First author Ahmad Ghazal)

Ahmad Ghazal is the main developer of DataGenX, affiliated with PingCAP, Inc., in Sunnyvale, CA, USA. PingCAP is the developer of the distributed SQL database TiDB, and DataGenX is an internal tool that has been released as part of the TiDB toolchain

The paper introduces DataGenX, a schema-driven, open-source tool for generating high-fidelity, scalable synthetic relational data. It addresses the common scenario where database developers need realistic data for simulation and benchmarking but only have the schema and limited statistics.

DataGenX operates by embedding declarative data generation logic within structured comments in standard SQL DDL. These annotations specify a comprehensive set of rules:

- Statistical Fidelity: Defines column distributions using functions (e.g., rand.zipf), conditional logic (CASE WHEN) to simulate value histograms, and explicit constraints on the Number of Distinct Values (NDV) to align with query optimizer statistics.
- Relational Integrity: Maintains foreign key relationships using shared variables and supports hierarchical generation (e.g., one parent row generating multiple child rows) for complex schemas like TPC-C.

The tool uses a highly parallelized, multi-threaded execution model to scale row generation efficiently across all available CPU cores, minimizing synchronization overhead.

To validate its accuracy, DataGenX was used to re-implement and generate data for the TPC-C benchmark. The resulting dataset achieved 99% fidelity against the official data generator on structural and statistical metrics. Furthermore, running the TPC-C workload on the generated data yielded comparable transaction throughput (TPM-C), confirming its suitability for rigorous performance evaluation.

Tabular Data Augmentation for Database Scalability Testing (First Author: Taro Fujimoto)

Taro Fujimoto is affiliated with Hitachi, Ltd., in Kanagawa, Japan. His paper is closely tied to commercial product development.

The paper addresses the challenge of performing database scalability testing when full-scale, real-world datasets are unavailable due to privacy concerns, such as with medical insurance claims. The solution is a hybrid Generative Adversarial Network (GAN)-based tabular data augmentation method to create synthetic data that preserves the original dataset's statistical and structural fidelity.

The proposed method combines two models: RC-TGAN (Relational Conditional Tabular GAN) is used to generate the join key attributes (claim_id) to maintain inter-table relational structure (1:N relationships). CTGAN (Conditional Tabular GAN) generates the non-key attribute columns, showing superior performance in accurately reproducing the often highly skewed distributions of categorical variables compared to the RC-TGAN-only approach.

In a case study involving Japanese medical insurance claims data, the synthetic datasets were generated at small (100K), middle (1M), and large (10M) claim scales. Scalability tests conducted on a commercial column-store database engine demonstrated:

- High Fidelity: The hybrid approach achieved high similarity (low Jensen-Shannon Divergence, JSD) in reproducing both relational structure and categorical variable distributions, which are crucial for affecting query selectivity and aggregation results.
- Scalability Validation: The execution time for the 10-query benchmark set consistently increased with the data scale, confirming the synthetic data's validity for practical database benchmarking.

System/Hardware Optimization and Sustainability (1 Paper):

DiStash: A Disaggregated Multi-Stash Transactional Key-Value Store (First author: Yiming Gao)

Yiming Gao is a researcher at the University of Southern California (USC), located in Los Angeles, CA, USA. This research was conducted collaboratively with engineers from eBay Inc., providing an academic perspective partnered with real-world enterprise application context.

The paper presents DiStash, a disaggregated transactional key-value (KV) store designed to manage data across multiple pools of diverse storage media, called "stashes" (e.g., DRAM, SSD, HDD, or NVM). The core innovation is enabling an application to use a single, ACID-compliant transaction type to read and write copies of data across these multiple stash pools. This single-transaction approach prevents complex race conditions that often arise when using different storage managers for different pools (e.g., a cache manager for DRAM and a separate DBMS for SSD), thereby ensuring data consistency and preserving isolation and durability properties. A stash pool's operation can be configured as either ephemeral (volatile storage for caching, evicting data when full) or durable (persistent storage with fixed capacity). The arrangement can be either inclusive (replicated) or exclusive (tiered).

DiStash is implemented by extending FoundationDB. In an evaluation using eBay's production workload, separating graph data (durable on SSDs) from query results (ephemeral on DRAMs) resulted in approximately a 10% performance improvement. This speedup occurs because the DRAM cache reduces SSD utilization, freeing up resources for other queries. DiStash was also shown to tolerate stash failures, using its Data Distributor to aggressively reconstruct replicas.

In addition to the peer reviewed regular papers we had two invited papers and one panel:

HammerDB: From Personal Project to Open Benchmarking Standard (Authors: Steve Shaw and Andy Bond)

Steve Shaw originated the HammerDB project (then Hammerora) as a personal open-source venture while employed at Intel. He helped establish the core architectural decision to use Tcl for its threading capabilities and implemented the first TPC-C workload for Oracle. He wrote the first published article featuring the Hammerora workload in the UK Oracle User Group magazine Oracle Scene in 2004. He is a published author, co-authoring the book Pro Oracle Database 10g RAC on Linux (2006). Following the project's adoption by the TPC, he officially joined the TPC-OSS committee to guide the project. He later founded HammerDB Ltd to manage the project full-time and led the significant upgrade to Tcl 9.0 for the v5.0 release

Andy Bond is affiliated with Red Hat Inc. Red Hat played a historic role in the benchmark landscape: a combination of Red Hat Enterprise Linux and the IBM DB2 database was the first x86-based system to exceed one million transactions per minute (tpmC) in 2009. He is a member of the TPC-OSS subcommittee, the group that sponsored and promotes HammerDB as a fair use implementation of TPC workloads.

The paper, "HammerDB: From Personal Project to Open Benchmarking Standard," recounts the journey of HammerDB, a leading open-source tool for database benchmarking, from its inception as a personal project to its current status as a mature industry solution. Widely adopted across enterprise, cloud, and academic environments, the tool is used to evaluate transactional and analytic database performance.

Originally called Hammerora, the tool was developed starting in 2003 by Steve Shaw for testing the Oracle database on Linux systems, particularly Oracle Real Application Clusters (RAC). The fundamental architectural choice was the Tcl/Tk language. Tcl was chosen for its lightweight footprint, cross-platform compatibility, and especially its robust, highly scalable native threading model, which allowed the tool to generate massive parallel workloads, unlike the concurrency-limited Python of the time.

Early implementations featured a TPC-C derived workload for transactional systems. A core design principle involved a simplified, two-stage methodology: automated, one-click schema creation and loading, followed by the workload generation using Virtual Users (threads). As adoption grew, especially after a featured article in the UK Oracle User Group magazine Oracle Scene in 2004, the tool added the TPC-H analytic workload and extended support to databases like MySQL and Microsoft SQL Server. This broadening of focus led to the software being renamed HammerDB.

The project introduced the NOPM (New Orders Per Minute) metric to enable performance comparison across different database engines without infringing on the TPC's trademarked "tpmC" terminology.

A major functional upgrade came with HammerDB v3.0, which added a Command-Line Interface (CLI), enabling automation in increasingly adopted cloud environments.

The significance of HammerDB was formally recognized in 2018 when the Transaction Processing Performance Council (TPC) created the Open Source Software Committee (TPC-OSS) and chose to sponsor, curate, and promote HammerDB as a model for fair use implementation of TPC workloads. This collaboration led to hosting the project on the TPC's GitHub repository. In v4.0, the benchmark workloads were renamed to TPROC-C and TPROC-H for compliance with TPC copyright policies. Further advancements supported cloud-native testing through features like Docker integration, a persistent SQLite web service for visualization, and a Python extension to the CLI.

The paper highlights the newest advancement in HammerDB v5.0 (released in spring 2025), which leverages the Tcl 9.0 release to enable a self-contained, single-file executable package for simplified, native installation on Linux and Windows. This marked the realization of the initial vision for a modern, portable, and easy-to-use benchmarking tool. The ongoing collaboration with the TPC and community aims to further integrate the tool into CI pipelines and explore submission of open-source benchmarks.

Next Generative AI Benchmark Development
(Authors: Hamesh Patel, Nirmala Sundararajan, Nicholas Wakou, Paul Cao, and David Schmidt)

Hamesh Patel represents Intel Corporation in this initiative. As a contributor from a major silicon and system architecture provider, his focus is on ensuring the benchmark accurately captures the performance characteristics and scalability of GenAI workloads across various hardware stacks. His involvement is critical to defining the metrics, such as tokens per second and throughput under concurrency, which are essential for evaluating large language models (LLMs) on high-performance infrastructure.

Nirmala Sundararajan works for Dell Technologies out of Austin, Texas, USA. Her participation is crucial for defining how the new TPC GenAI standard will measure performance and cost efficiency in production environments. Her work focuses on addressing the complexities of real-world deployments, including mixed workloads, concurrency, and resource management, which are often ignored by existing benchmarks.

Nicholas Wakou works for Dell Technologies out of Austin, Texas, USA. His role involves addressing the unique demands of modern GenAI workloads, which differ significantly from traditional inference tasks. His contributions help shape the framework to evaluate emerging model types—such as multi-modal architectures and Retrieval-Augmented Generation (RAG) systems—that require specific computational patterns and latency profiles.

Paul Cao works for HPE. He contributes expertise on full-stack evaluation and end-to-end pipeline benchmarking. His perspective is key to defining the degrees of freedom allowed to test sponsors, such as varying the model layout, inference engine, vector database, and document store, while ensuring the results remain reproducible and auditable in line with TPC's principles.

David Schmidt is affiliated with Red Hat in Raleigh, North Carolina, USA. Representing a leading enterprise software provider, his focus is on the system and framework considerations, particularly how the benchmark accounts for the entire inference stack, including model layout, sharding, KV cache management, and software optimizations. His contributions help ensure the resulting benchmark is architecture-neutral and tests a wide range of models across diverse software stacks.

The paper addresses the urgent need for a new industry-standard benchmark for Generative AI (GenAI) workloads. The existing benchmarks do not adequately capture the unique computational and data-centric characteristics of models like Large Language Models (LLMs), diffusion models, and multimodal architectures. Building on the foundation of TPCx-AI, the Transaction Processing Performance Council (TPC) is developing this new benchmark to reflect real-world, production-scale GenAI scenarios.

The new GenAI benchmark will focus on generative tasks such as text generation and summarization, code generation, and multi-modal tasks, specifically excluding object detection and classification. It is designed to evaluate the full end-to-end pipeline, encompassing data ingestion, inference, post-processing, and quality assessment. In line with TPC's established principles, it will measure three primary metrics: Performance, Price/Performance, and Hardware and Software Availability. The final result will be measured in terms of tokens/sec per dollar and efficiency under multi-user workloads. To ensure the benchmark genuinely drives performance improvements, it is being designed to be resistant to Goodhart's Law by incorporating task-specific metrics that align with production criteria, rather than only easily optimized targets like raw throughput or accuracy.

Understanding the responsiveness and scalability of GenAI systems is critical. Key performance metrics include Time-to-First-Token (TTFT), which is vital for interactive applications, and Request Latency, which captures the end-to-end response time. System throughput is measured by Tokens per Second and Throughput Under Concurrency, which assesses scalability under simultaneous requests.

Beyond speed, the benchmark will integrate Cost and Efficiency Metrics. As models grow in complexity, financial and environmental costs are central concerns. Tracking Energy Consumption across training and deployment is essential for sustainability. Furthermore, Cost per Query quantifies the financial overhead for each model invocation, enabling comparative analysis and informing deployment strategies.

Evaluating the output quality of GenAI models requires a multifaceted approach due to the non-deterministic nature of LLMs.

- Quantitative and Task-Based Metrics: The evaluation will use structured benchmarks like MMLU (Massive Multitask Language Understanding) to test cross-domain generalization and HELM (Holistic Evaluation of Language Models) to assess reliability, robustness, and ethical behavior. Traditional probability-based metrics, such as Perplexity and KL Divergence, will provide statistical insights but are noted for their limited correlation with human-perceived quality in creative tasks.

- Subjectivity and Human Review: Given the limitations of automated metrics, Human-in-the-Loop (HITL) Assessment is a critical component. Experts use rubric-based scoring to judge subjective dimensions like creativity, coherence, and factual accuracy, especially in high-stakes fields like medicine or law where model hallucinations pose significant risks.

A major hurdle is reproducibility. Small changes in prompt wording, context length, model versions, or even the underlying hardware can lead to inconsistent results. The complexity of the entire inference stack, including model sharding, KV cache management, and vector databases, further complicates the measurement of performance and quality.

LLM inference involves two distinct phases: prefill, where the prompt context is built (often compute-bound), and decoding, where new text is generated one token at a time (often memory-bound). The system's memory demands are driven by the large, static model weights and the highly memory-intensive key-value (KV) cache.

Optimizations are critical for improving the compute-to-communication ratio. These include quantizing weights to use smaller data types, virtualizing memory using techniques like PagedAttention to reduce fragmentation, and using I/O-aware operations like FlashAttention to limit memory accesses. Furthermore, techniques like speculative decoding expose additional parallelism by using a smaller draft model to accelerate the decoding phase, dramatically improving latency without sacrificing quality. Real-world production systems also require complex application logic for routing requests, orchestrating pipelines (e.g., in summarization), and providing context through Retrieval-Augmented Generation (RAG).

The TPCx-AI subcommittee is actively defining the specific parameters, constraints, and necessary disclosures to create a fair, robust, and enduring benchmark that can adapt to the rapid evolution of multimodal AI.

Panel: Benchmarking Agentic AI Systems - A Narrative Summary

(Authors: Ajay Dholakia, Sachin Gopal Wani, David Ellison, Miro Hodak, Debojyoti Dutta, Shishir Nagaraja)

Ajay Dholakia is a Principal Engineer, Master Inventor, AI Leader, and Chief Technologist for Software & Solutions Development with Lenovo ISG. His focus includes solution architectures in AI/ML, Generative AI, Data Analytics, Edge Computing, and Blockchain. His career, spanning over 30 years with Lenovo and IBM, has involved leading diverse projects from research to product development and technical strategy. He holds more than 60 patents and has authored over 60 technical publications, including a book. He earned a PhD in Electrical and Computer Engineering from N.C. State University and an MBA from Henley Business School.

Sachin Gopal Wani is an AI Data Scientist at Lenovo. He works on end-to-end Machine Learning (ML) applications for customers and is involved in developing the NewTalk AI framework. He graduated from Rutgers University as a gold medalist specializing in Machine Learning and secured the J.N. Tata Scholarship. He has authored documents including a Lenovo LLM Sizing Guide and one on the Total Cost of Ownership of Generative AI.

David Ellison is the Chief Data Scientist and Director of AI Engineering for Lenovo ISG. He leads a team through Lenovo's US and European AI Discover Centers that uses cutting-edge AI techniques to deliver solutions for external customers and support the overall AI strategy for the Worldwide Infrastructure Solutions Group. Before joining Lenovo, he ran an international scientific analysis and equipment company and worked as a Data Scientist for the U.S. Postal Service. He holds a PhD in Biomedical Engineering from Johns Hopkins University and has numerous publications in top-tier journals, including two in the Proceedings of the National Academy of the Sciences.

Miro Hodak is a Principal Member of Technical Staff at AMD, where his work focuses on AI performance and benchmarking. He previously served as an AI Architect at Lenovo and was a professor in physics at North Carolina State University. Since 2020, he has been actively involved with MLPerf and MLCommons, contributing to multiple MLPerf benchmarks and serving as co-chair of the MLPerf Inference Working Group since 2023. He has authored peer-reviewed publications across fields including AI, computer science, materials science, physics, and biochemistry.

Debojyoti (Debo) Dutta is the Chief AI Officer at Nutanix, where he drives AI initiatives and develops generative AI products, including Nutanix Enterprise AI. He previously held the position of Vice President of Engineering at Nutanix, overseeing multi-cloud AI operations, governance, and data engineering for SaaS products. He is a Board Advisor and Independent Observer for MLCommons and a former Distinguished Engineer at Cisco Systems. He holds a PhD in Computer Science from the University of Southern California and a B. Tech in Computer Science & Engineering from the Indian Institute of Technology, Kharagpur.

Shishir Nagaraja is a Professor of Cybersecurity in the School of Computing at Newcastle University. His research interests focus on network security and privacy. His research spans various topics in cybersecurity, with numerous peer-reviewed publications covering areas like software supply chain security, network traffic analysis, and security in medical and control systems.

Raj Ranjan (also known as Rajiv Ranjan) is an Australian-British computer scientist and University Chair Professor for the Internet of Things (IoT) research in the School of Computing at Newcastle University. He is known for his research in Distributed Systems, including Cloud Computing, Big Data, and the Internet of Things. He is the director of the Networked and Ubiquitous Systems Engineering (NUSE) Group and the Academic Director of the School of Computing. He is a highly cited author and has secured over £12 Million GBP in competitive research grants.

The paper addresses the crucial and evolving subject of Benchmarking Agentic AI Systems , which represents a pivotal shift from generative AI. An Agentic AI system is defined by its ability to interact with an environment to achieve a goal with a degree of autonomy. A collection of AI Agents are often stitched together to design and deploy these systems for enterprise-level workflows.

Benchmarking Agentic AI necessitates a fundamental departure from the benchmarks traditionally used for large language models (LLMs). While generative model metrics focused on the quality of a static output, agentic evaluation must assess a dynamic, interactive process. An agent's performance is defined by the entire trajectory of actions it takes, including its ability to reason through multi-step problems, use external tools, and adapt to unforeseen circumstances in a changing environment.

The Agentic AI process itself is non-linear and stateful, involving planning, executing actions (with tools), and evaluating the result in an iterative loop. This flow contrasts sharply with traditional AI benchmarking, which is typically a single model providing a prediction for a given input. Proper benchmarking, therefore, must account for the AI model executing the planning step, the tools that execute actions, and the AI model evaluating the result, along with the non-linear flow.

Early efforts to standardize evaluation led to general, multi-dimensional benchmarks like AgentBench, which assesses an LLM's general reasoning and decision-making abilities across eight distinct environments. For practical application in enterprise settings, highly specialized, task-oriented benchmarks have emerged. For instance, SWE-bench presents agents with real-world problems sourced from GitHub issues, requiring them to generate a code patch that passes a rigorous execution-based "fail-to-pass" criterion. ITBench addresses live system operations, evaluating an agent's ability to act as a Site Reliability Engineer (SRE) or analyst, using business-relevant metrics like Mean Time to Repair (MTTR).

Beyond accuracy, throughput, and latency, agentic AI demands additional metrics that reflect real-world viability. Process-oriented metrics are used to diagnose failure modes within long traces. For example, AgentQuest introduces "progress rate" and "repetition rate" to identify failure modes like looping or stagnation. Furthermore, benchmarks like TRAIL evaluate an agent's ability to debug and localize errors in its own long execution traces, which is critical for sustained autonomy.

The autonomy and complexity of agentic AI make it difficult to govern. A robust governance strategy, which is a prerequisite for responsible deployment, involves adhering to international standards and ensuring comprehensive auditability. Tools like the NIST AI Risk Management Framework (AI RMF) and the international standard ISO/IEC 42001 are foundational for managing risks and ensuring responsible use. A central challenge is the "accountability gap" created by opaque decision-making. This necessitates building systems with immutable audit trails that log every agent action and its reasoning path for traceability.

Security emerges as a paramount concern; where agents autonomously take decisions, security failures will manifest as safety failures. This creates a higher legal onus on companies. The responsibility for decision quality lies with the AI agent itself. Agentic AI faces security challenges like Goal-integrity (where mimicry attacks can corrupt decision boundaries) and Pattern integrity (where agents learn harmful patterns from adversarial agents).

Finally, practical deployment requires assessing Cost per task. Benchmarks are increasingly reporting resource consumption—such as tokens, external tool calls, compute time, and energy—as part of a normalized cost. Responsible operation also requires clear human-in-the-loop safeguards, asking whether an agent knows when to escalate and respects policy boundaries for high-risk actions.

The paper concludes that benchmarks must evolve from single-model accuracy toward system-level assessment, integrating functional (performance, correctness) and non-functional (governance, security, cost) requirements to guide the field toward effective, efficient, and trustworthy systems.

TPCTC 2025 Organization

General Chairs and Contacts

Raghunath Nambiar, AMD, USA, raghu.nambiar@amd.com

Meikel Poess, Oracle, USA, meikel.poess@oracle.com

Publication Chair

Rodrigo D. Escobar, Univ. Texas at San Antonio, USA

Program Committee

Ahmad Ghazal, PingCAP, USA

Ajay Dholakia, Lenovo, USA

Andrew Bond, Red Hat, USA

Hans-Arno Jacobsen, University of Toronto, Canada

Hanumath Rao Maduri, Workday, USA

Harry Le, University of Houston, USA

John Poelman, IBM, USA

Karthik Krishna, InfobellIT Solutions, India

Klaus-Dieter Lange, Hewlett Packard Enterprise, USA

Michael Brey, Oracle, USA

Miro Hodak, AMD, USA

Nicholas Wakou, Dell, USA

Paul Cao, Hewlett Packard Enterprise, USA

Pratyush Agnihotri, Technical University of Darmstadt, Germany

Rodrigo D. Escobar, Univ. Texas at San Antonio, USA

Shahram Ghandeharizadeh, University of Southern California, USA

Tariq Magdon-Ismail, VMware, USA

Tilmann Rabl, Hasso Plattner Institute, Germany