OTTO AI Conference

‹MAIN›

Session 3.0

28.11.2024

OTTO-Campus

Hamburg & Digital

For AI Developers

Blue Stream

(Eventfläche)

 **Sarah Gritzka (OSP)**  **Thanh Quang Nguyen (OTTO)**

Gen-AI in Customer Service / 10:05 – 10:45 Uhr

GenAI presents significant potential in transforming OTTO customer service. By handling first-level emails, it eliminates the need for human intervention in repetitive tasks, enabling our agents to focus on more complex challenges. Additionally, GenAI supports agents by summarizing emails, streamlining their workflow. This innovation not only improves efficiency but also reduces costs and human contact needs. However, challenges such as hallucination, data privacy issues, and the requirement for continuous updates accompany these benefits. In our talk, we will explore these challenges and share our strategies for overcoming them, ensuring that GenAI enhances our customer service operations effectively.

Pink Stream

(Konferenz 1)

 **Steven Mc Auley (Google)**  **Juri Wiens (REWE Digital)**

Agentic Systems: Von verborgenen Fallstricken und wertvollen Einsichten / 10:05 – 10:45 Uhr

LLM-basierte agentische Systeme eröffnen faszinierende Möglichkeiten – aber auch Herausforderungen, die oft erst beim Entwickeln sichtbar werden. Dieser Vortrag beleuchtet die „Unknown-Unknowns“ und zeigt anhand von Architekturprinzipien, Kommunikationsmustern und Workflow-Strategien, wie verborgene Stolpersteine zu wertvollem Wissen werden. Die Teilnehmer erfahren, welche Einsichten in der Entwicklung skalierbarer, agentischer Systeme entscheidend sind – und was man sich gerne schon früher gewünscht hätte zu wissen.

Blue Stream

(Eventfläche)

 **Marcel Jahnke &**  **Kevin Ness (TK)**

RAG in der Praxis: Erfahrungen aus der Techniker Krankenkasse / 11:10 – 11:50 Uhr

In der gesetzlichen Krankenversicherung (TK) bietet die Retrieval-Augmented-Generation (RAG) große Potenziale für die Verbesserung von Informationsabruf und -verarbeitung. In diesem Vortrag teilen wir unsere Erfahrungen mit der Implementierung von RAG in der TK und diskutieren die Herausforderungen und Lösungen.

Wir werden auf die wichtigsten Aspekte der RAG-Implementierung eingehen, darunter die Integration von Large Language Models, die Entwicklung von Retrieval-Strategien und Prompts, die Implementierung von Feedback-Mechanismen sowie die Rolle von GPU-on-Prem-Frameworks und Embedding-Datenbanken.

Pink Stream

(Konferenz 1)

 **Masha Stroganova (Microsoft)**

Leveraging agentic patterns for solving complex AI problems / 11:10 – 11:50 Uhr

In this presentation, we will explore the use of agent-based architecture patterns for intelligent applications and demonstrate how combining individual large language models can solve more complex problems with the help of artificial intelligence. Multi-agent systems become interesting for scenarios where the technical complexity of a multi-stage process can no longer be reliably and efficiently handled by a single model or prompt. Desired results can be achieved through feedback and review cycles with other models. We will present the fundamental concepts, suitable problems, a few practical examples, demos, and architectures to discuss the exciting possibilities of the next generation of AI applications.

Blue Stream

(Eventfläche)

Jan Strich (Uni Hamburg)

RAG - State-of-the-art approaches and evaluation opportunities / 11:55 – 12:20 Uhr

Retrieval-augmented generation (RAG) improves LLMs by integrating external data, and this presentation explores SOTA techniques to optimize retrieval accuracy and relevance. Starting from simple RAG pipelines using simple query matching with cosine similarity, the field has shifted to more complex methods that further improve RAG. Key innovations include intelligent reranking with LLM-based scoring, GraphRAG, RAPTOR, and ensemble and multimodal retrieval for richer, more diverse answers. In addition, one of the biggest challenges is the evaluation of RAG systems due to the complexity of the system. While standard benchmarks such as BLEU and ROUGE are commonly used to evaluate answer quality, the RAGAS framework goes beyond this by including metrics such as answer fidelity, context precision, and answer relevance. These metrics enable a more comprehensive evaluation of the entire RAG system and ensure both retrieval accuracy and context adaptation.

Blue Stream

(Eventfläche)

Florian Schneider (Uni Hamburg)

Bridging Modalities: Cross-Modal Retrieval and Vision Language Models for Multimodal AI Technology / 12:20 - 12:45 Uhr

Cross-Modal Retrieval and Vision Language Models for Multimodal AI Technology Advancements in multimodal AI are redefining how we process and interpret diverse forms of data. This talk explores cross-modal transformer encoder models, which compute dense vector representations of multimodal data such as images and text within the same vector space. These models form the foundation of cross-modal retrieval systems. By aligning these vector embeddings, cross-modal encoder models enable efficient similarity searches across different data types. Expanding on this, we delve into the role of multimodal Large Language Models (LLMs) or Vision Language Models (VLMs) used for Multimodal Retrieval-Augmented Generation (MuRAG). These models facilitate detailed inquiries about images and allow for ranking them according to specific search criteria, enhancing the ability to extract nuanced information from visual content. The talk highlights the architectural differences and similarities between cross-modal transformer encoders and VLMs, providing insights into how VLMs are constructed and operate. Through a detailed examination of their underlying mechanisms, attendees will gain a deeper understanding of the strengths and applications of each model archetype and its role in MuRAG systems. Concluding with a brief live demonstration of a simple MuRAG system in the realm of academic online collections, the presentation showcases a practical application of these technologies, illustrating their potential impact on future multimodal AI systems.

Pink Stream

(Konferenz 1)

Prabesh Dhakal (bonprix)

Ranking Items Based on User Behavior: Lessons Learned Along the Way / 11:55 bis 12:35 Uhr

In this talk, we will explore Behavior Cluster Ranking, an approach to ranking fashion items based on users‘ behavior rather than static factors like age or aggregated item-level purchase history. By focusing on the actions of the users - such as adding items to the cart or changing preferences - we can better understand and respond to users‘ current interests. We will also highlight the challenges we faced and the lessons we learned along the way.

Blue Stream

(Eventfläche)

Johannes Langer &

Michelle Mei-Li Pfister (AWS)

10 Tips for building production ready RAG applications / 14:00 - 14:40 Uhr

Developing robust, scalable, and production-ready Retrieval Augmented Generation (RAG) applications is challenging, but with the right strategies and best practices, you can ensure your applications are reliable, and secure. In this session, we‘ll explore 10 essential tips that will help you navigate the challenges of building production-ready RAG applications.
By the end of this session, you‘ll have a comprehensive understanding of the key strategies and best practices for building production-ready RAG applications. Join us to elevate your RAG application development skills and create solutions that thrive in real-world production environments.

Pink Stream

(Konferenz 1)

Dr. Felix Rochau (ABOUT YOU)

Explainability in Multi-Touch Attribution: Moving beyond rule-based models / 14:00 - 14:40 Uhr

In the competitive world of E-Commerce, measuring and evaluating channel performance is essential for optimizing marketing ROI. This talk examines the shift from traditional rule-based attribution models to data-driven Multi-Touch Attribution (MTA) approaches, highlighting the challenge of ensuring Explainability in more complex models.

The presentation will introduce several MTA models that have been tested, revealing that they often produce varying results that are not always easily interpretable. Incrementality Testing is then introduced as a crucial method for evaluating these models, helping to assess their accuracy and consistency. The talk will conclude with insights into a Unified Marketing Measurement approach, which integrates a Bayesian Marketing Mix Model to calibrate traditional approaches using experimental priors for a more holistic strategy.

Blue Stream

(Eventfläche)

Dr. Daniel Beck (HHLA)

GenAI in der intermodalen Logistik / 14:45 - 15:25 Uhr

Effiziente und skalierbare Automatisierungslösungen sind in der maritimen und intermodalen Logistik von zentraler Bedeutung. Die Hamburger Hafen und Logistik AG (HHLA) verfolgt Ansätze, die cloudbasierte Backend-Prozesse mit Generative AI (GenAI) und klassischen Prognosemodellen kombinieren. Multiagentenbasierte Systeme spielen dabei eine Schlüsselrolle: Intelligente KI-Agenten übernehmen unterschiedliche Aufgaben, arbeiten eigenständig und koordinieren sich in Echtzeit, um Prozesse wie die Bestellabwicklung, die dynamische Zuordnung von Schiffsladungsdaten und die effiziente Verarbeitung von Zolldokumenten zu optimieren. Diese Systeme sind darauf ausgelegt, nahtlos mit ERP-, Terminal Operating- und Mail-Systemen zu interagieren, um eine flexible und leistungsstarke Automatisierungslösung zu bieten. Der Vortrag zeigt anhand verschiedener Anwendungsfälle, wie diese Technologien dazu beitragen können, die Effizienz und Flexibilität in den Logistikprozessen der HHLA zu steigern.

Pink Stream

(Konferenz 1)

 **Henrike Meyer &**  **Theresa Wohlsen &**  **Lukas Janssen (Otto Group / data.works)**

Image Generation 101: Understanding the Fundamentals / 14:45 - 15:25 Uhr

Join us for an introduction to the 101‘s of image generation with generative AI._ Henrike and Theresa from Team TIGA introduce you to the big world of image generation with stable diffusion models! Learn how the technology works and discover the differences between commercial and open source approaches with their strengths and weaknesses.

Blue Stream

(Eventfläche)

 **Dr. Dirk-Sören Luehmann (Kühne+Nagel)**

Artificial Intelligence in Logistics - Leveraging Digital Twins for AI-driven Price Optimization / 15:30 - 16:10 Uhr

Intelligent pricing in logistics is tailored to customers and market conditions. We use artificial intelligence (AI) to generate optimized prices and a digital twin to simulate strategic decisions and KPIs. Our AI models face various challenges, including sparse data, configurable products, and local market conditions. Our ultimate goals are customer-specific price elasticities and AI predictions explainable to humans. In my presentation, I will explain how we use AI for price optimization and its integration into Kuehne+Nagel‘s daily operations.

Pink Stream

(Konferenz 1)

Christoph Eiwen &

Marlon Kaulich (Rossmann)

Large AI Projects with a Small Team: Open-Source at ROSSMANN / 15:30 - 16:10 Uhr

The Prompt Engineers at ROSSMANN demonstrate how combining open-source projects with custom adaptations and both internal and external expertise can lead to significant successes. The flexible „Make and Buy“ approach enables the development of powerful solutions like RossmannGPT and the Store Copilot, even with a small AI team. RossmannGPT integrates classic ChatGPT features with meaningful functionality enhancements into a user-friendly interface. The Store Copilot supports store employees with an advanced RAG system for quick and precise information retrieval. With minimal effort, maximum benefit is derived from AI technologies, shaping the future of AI-assisted support within the company.

Blue Stream

(Eventfläche)

Dr. Marina Runge (INWT)

Predictive LLMs: Prognose von Gebrauchtwagenpreisen mit LLMs vs. XGBoost / 16:30 - 17:10 Uhr

Large Language Models (LLMs) wie GPTs (Generative Pre-trained Transformers) haben sich als leistungsfähige Werkzeuge zur Verarbeitung und Generierung natürlicher Sprache etabliert. In diesem Vortrag wird eine bisher eher untypische Anwendung dieser Modelle vorgestellt: die Vorhersage einer metrischen Zielvariablen. Diese Art der Anwendung geht über typische textbasierte Aufgaben hinaus und eröffnet neue Möglichkeiten im Bereich der datengetriebenen Vorhersage. Durch ihre Fähigkeit, sowohl strukturierte als auch unstrukturierte Daten - wie Freitext - zu verarbeiten, bieten LLMs besondere Vorteile bei der Merkmals-generierung und der Erkennung komplexer Muster. Können diese Eigenschaften genutzt werden, um die Vorhersagegenauigkeit bei numerischen Zielvariablen zu erhöhen? Und welchen Mehrwert bieten LLMs im Vergleich zu etablierten Methoden wie XGBoost? Verschiedene Anwendungsfälle werden untersucht, darunter die Arbeit mit rein tabellarischen Daten, die Kombination von tabellarischen Daten und Freitext sowie die Generierung von Merkmalen aus Textdaten. Neben der Verwendung von GPT-3.5 von OpenAI werden auch Open-Source-LLMs betrachtet, die durch lokale Verarbeitung datenschutzfreundliche Alternativen bieten. Am Beispiel der Vorhersage von Gebrauchtwagenpreisen wird demonstriert, wie LLMs zur Verbesserung der Vorhersagequalität von metrischen Zielvariablen beitragen können. Der Vortrag gibt praktische Einblicke in den Einsatz von LLMs für Predictive Analytics und beleuchtet deren Potenzial im Vergleich zu “klassischen” Machine-Learning-Verfahren wie XGBoost.

Pink Stream

(Konferenz 1)

Cara Watermann (Vodafone)

GenAI meets German Precision: Implementing RAG in a company setting / 16:30 - 17:10 Uhr

Regulations and corporate environment restrictions are a limiting factor when trying to turn GenAI POCs into running products in Germany. Together we explore the challenges of building GenAI tools in a German corporate environment, from architectural decisions to workers‘ council approvements to security and data privacy processes. We will present how we tackled those challenges and introduced an internal LLM-based Assistant for Vodafone Germany.