Die 95 Prozent sind ein Messproblem

Am zweiten GenAI Enterprise Day bei IBM in Zürich zeigten fünf Praktikerinnen und Praktiker, was KI, die sich auszahlt, von KI unterscheidet, die nur in der Demo überzeugt: ein Ausgangswert, ein gemessener Arbeitsablauf und eine Evaluation, die weiterläuft.



VERÖFFENTLICHT

26. November 2025

ONLINE LESEN

samuylov.ai · Die 95 Prozent sind ein Messproblem

Ein Projekt, das seinen Ausgangspunkt nie gemessen hat, kann keinen Ertrag nachweisen. Im November 2025 war das die nützlichere Lesart der 95-%-Schlagzeile.

Die Schlagzeile ging auf den Bericht *The GenAI Divide* von MIT NANDA zurück. Messbarkeit war das Thema des zweiten GenAI Enterprise Day, *GenAI Reality Check: Business Wins, Fails & the Metrics in Between*, am 26. November 2025 im IBM Auditorium in Zürich, Teil von GenAI Enterprise Hub. Denis eröffnete den Abend mit Lars Mallien von IBM Schweiz und moderierte das Abschlusspanel.

Fünf Geschichten

- **Im grossen Massstab gemessen.** Markus Danhel, IBM, über IBM als ersten eigenen Kunden – mit Produktivität, die in HR, IT-Support, Finanzen, Einkauf und Lieferkette gemessen wird.
- **In Etappen skaliert.** Katarina Granath, SIX, über drei Jahre vom ersten Bot bis zur Unternehmensplattform, mit Governance und AI Integration Leads direkt im Geschäft.
- **Zuerst messen.** Alexander Taboriskiy, Mentiora AI, darüber, ein System anhand geschäftlicher Anforderungen zu definieren, zu messen und zu verbessern.

- **Von Anfang an verantwortungsvoll.** Anne-Marie Buzatu, ICT4Peace Foundation, über Risikobewertung, Schutzmechanismen und die Fähigkeiten der Mitarbeitenden.
- **Stabil, wenn es darauf ankommt.** Manuel Flurin Hendry, Silicon Stories, darüber, Echtzeit-Inferenz während einer Live-Aufführung stabil zu halten.



Der Einführungsweg von SIX, 2023–2025.

Was das Messen verändert hat

Das deutlichste Beispiel kam von Mentiora. Auf einem geschäftlichen Evaluationsset stieg die Erfolgsquote mit denselben zugrunde liegenden Modellen von 31 % auf 91 %. Verändert hatte sich eine schriftliche Definition dessen, wie eine richtige Antwort aussieht – und ein kalibrierter Modell-Judge, der jede Antwort daran prüfte.



Die Lücke zwischen einer Demonstration und verlässlichem Einsatz.

Alle fünf kamen für das Abschlusspanel zurück, unter drei Regeln, die zu Beginn des Abends festgelegt worden waren: keine Verkaufspräsentationen, Fragen stellen, Erfahrungen teilen.



Das Abschlusspanel im IBM Auditorium.

Was das für Ihr Projekt bedeutet

Bevor Sie einen Arbeitsablauf verändern, messen Sie ihn: Volumen, Bearbeitungszeit, Fehlerquote und Prüfaufwand. Halten Sie fest, wie eine richtige Antwort aussieht, und prüfen Sie vor und nach der Einführung dagegen. Dann zeigt ein Pilotprojekt entweder seinen Wert – oder genau, wo es zu kurz greift.

Unser System zur Dokumentenklassifizierung hat diese Prüfung eingebaut: Jedes Dokument wird zweimal unabhängig klassifiziert, und ein KI-Judge entscheidet bei Abweichungen.

Sie wollen einen Assistenten oder Agenten evaluieren? Beginnen Sie beim Ausgangswert. Wir legen ihn mit Ihrem Team vor dem Bau fest und messen danach weiter. Siehe Workshops & Training, oder Kontakt aufnehmen.



Denis Samuylov
– Gründer & CEO

Vom Gründer geführt, vom Team umgesetzt – Denis leitet das Mandat und bleibt auch nach dem Pitch dabei.

contact@samuylov.ai +41 79 136 38 26 Erstgespräch buchen →