

masterclass

hybrid

1-day

Testing General Purpose AI (GPAI) applications

As GPAI models continue to improve, systems are increasingly touted as matching or exceeding expert performance in “problem solving”, “scientific reasoning” or “real-world software engineering.” These claims are based on common tests for different capabilities, known as “benchmarks”. Yet, there is often misconception of what benchmarks actually measure, what conclusions about a system we can actually make and how to evaluate what actually matters for a use case. Just as other technology-driven industries, like healthcare and aviation, AI benchmarking needs reliable standards. With best practices in their infancy, it can be daunting for professionals to navigate GPAI evaluation.



In this masterclass, we make accessible the most valuable insights from the field of GPAI benchmarking. The course covers:

- > GPAI under the AI Act and benchmarking for systems with systemic risk
- > Prominent industry benchmarks and adaptations to GPT-NL
- > Insights from Algorithm Audit's own benchmarking work
- > The design aspects that determine the quality of a benchmark, incl. construct validity

After attending, participants will be able to:

- > Grasp how benchmarks tend to work “under the hood”
- > Navigate common evaluation repositories and documentation
- > Assess the suitability of existing benchmarks for their own work



Tue Nov 3rd 10:00–17:00 CET



Hybrid and The Hague
Conference Centre (New Babylon)
Anna van Buerenplein 29, 2595 DA
Den Haag



€ 300 (in-person)
(including, lunch, drinks, refreshments)

For whom

Max. 25 professionals from private and public sector who regularly work with GPAI application, such as implementation of generative AI solutions in work processes, testing GPAI capabilities and/or working on AI policy.

Instructors



Jurriaan Parie

Director and board member
at Algorithm Audit



Kaloyan Filipov

AI Advisor and Validator at
Algorithm Audit

Program

Time	Topic	Instructor
09:30-10:00	Doors open	
10:00-11:15	Introduction General Purpose AI (GPAI) and testing GPAI applications	Jurriaan Parie
11:15-11:30	<i>Break – coffee, tea and refreshments</i>	
11:30-12:30	State-of-the-art concepts and developments in GPAI benchmarking	Kaloyan Filipov
12:30-13:30	<i>Lunch – catered</i>	
13:30-14:45	Case study from Dutch public sector	Jurriaan Parie
14:45-15:15	<i>Break – coffee, tea and refreshments</i>	
15:15-16:30	Hands-on exercises to build practical experience	Jurriaan, Kaloyan
16:30-17:30	<i>Drinks – catered</i>	