Recent progress on standardization efforts around Benchmarking AI in telecom networks

Antonio De Domenico, PhD
Principal Engineer
Huawei Technologies France
antonio.de.domenico@huawei.com

Contents

HUAWEI

Major Large Language Models (LLMs)

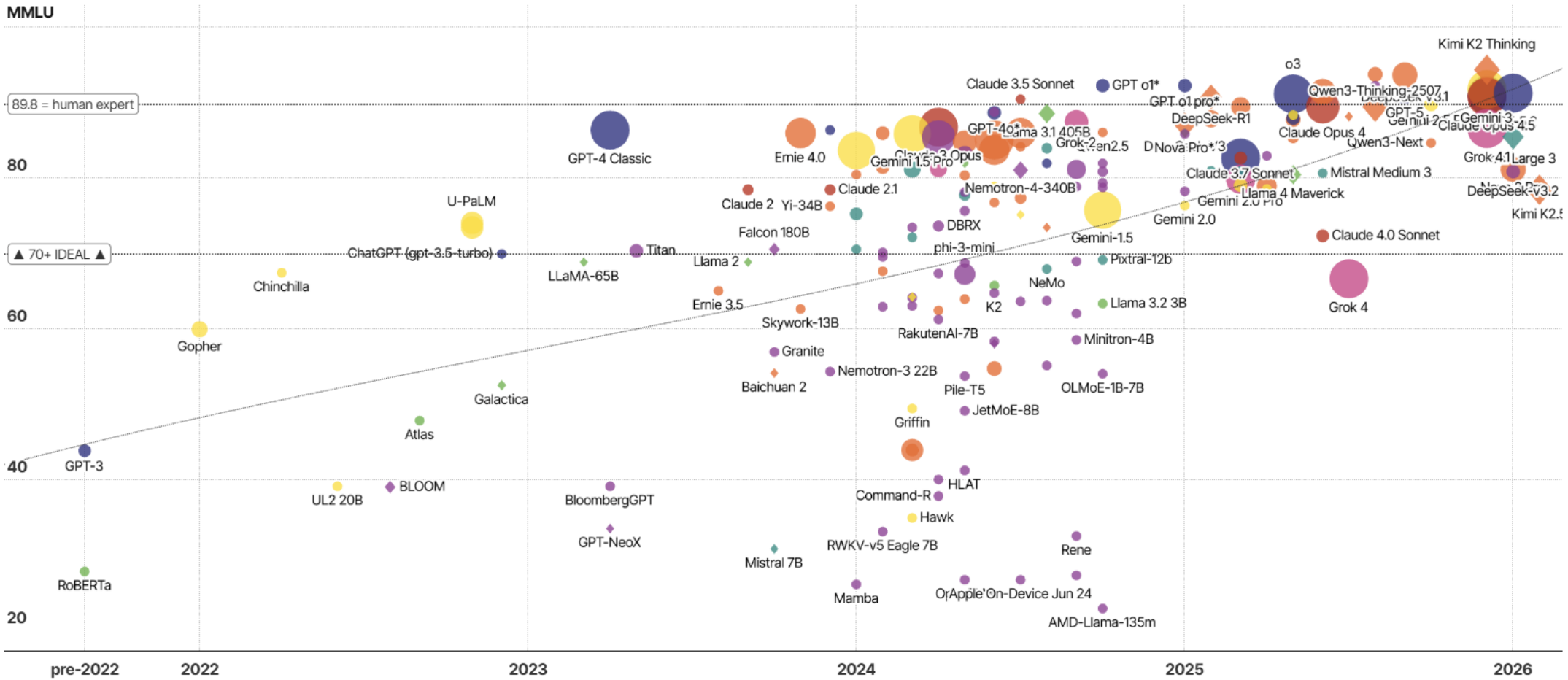
ranked by capabilities, sized by billion parameters used for training

Parameters (Bn) open access

CLICK LEGEND ITEMS TO FILTER

anthropic chinese google meta mistral openAI other xAI

search... show only: all



David McCandless, Tom Evans, Paul Barton

Informationisbeautiful // Jan 2026

MMLU = benchmark for measuring LLM capabilities

* = parameters undisclosed // source: LifeArchitect // data

Role of AI in telecom networks

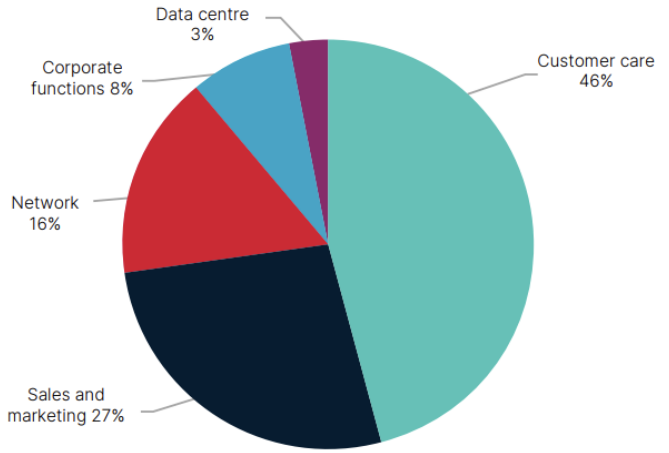
MNO activity continues to gravitate towards low-risk, easy wins where functions can be automated through (simple) AI agents

Customer care accounts for almost 50% of deployment

However 34% of total OpEx spent on network operations

Investments on network and data-centre reflects a desire to monetise inference and compute capabilities

Where is the AI sweet spot for telcos?
Percentage of telco AI deployments



Source: GSMA Intelligence

Telcos deploy AI in different parts of their business

Deployment area	Example use cases
Network	Maintenance AI-RAN Energy saving Performance
Product sales and marketing	Customer sales Application development Solution provision (e.g. operators working with a third party to deliver AI-enabled services to enterprise customers, such as GPUaaS)
Customer care	Chatbots Spam filters User experience
Corporate functions	Employee development Operational support (e.g. finance, IT) Energy saving
Data centre	OSS/BSS

Source: GSMA Intelligence

-GSMA Intelligence, Telco AI: State of the Market, Q1 2026
https://www.gsmaintelligence.com/research/telco-ai-state-of-the-market-q1-2026

Benchmarking AI in Telecom tasks

The smaller the model size, the less energy the model requires to generate one token

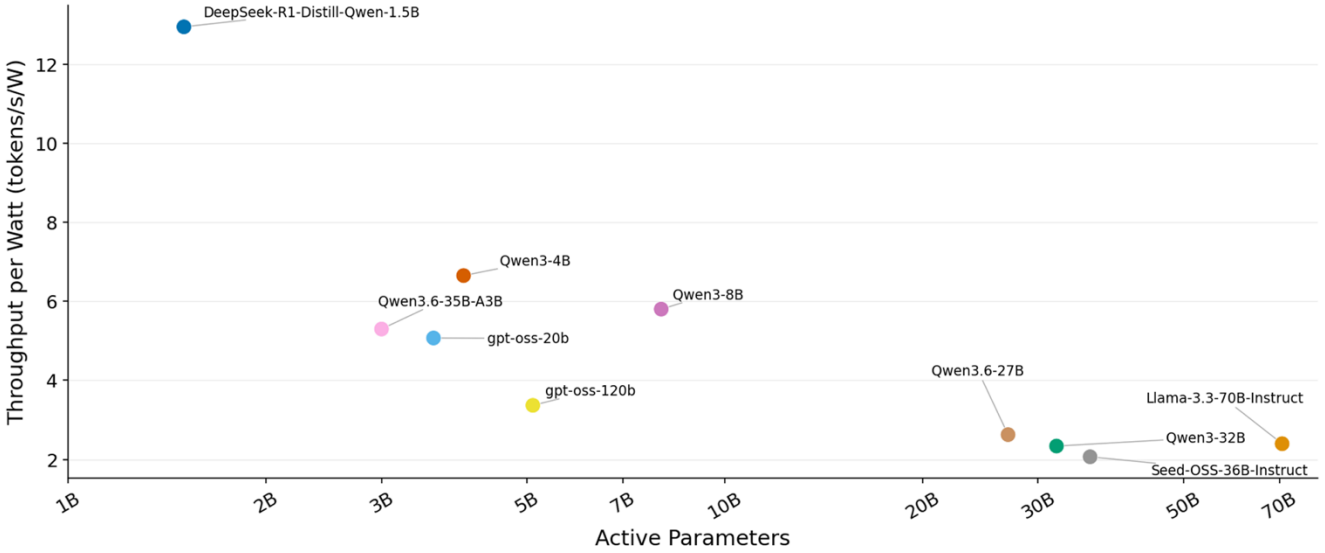
Small models may not be strong enough for telecom tasks

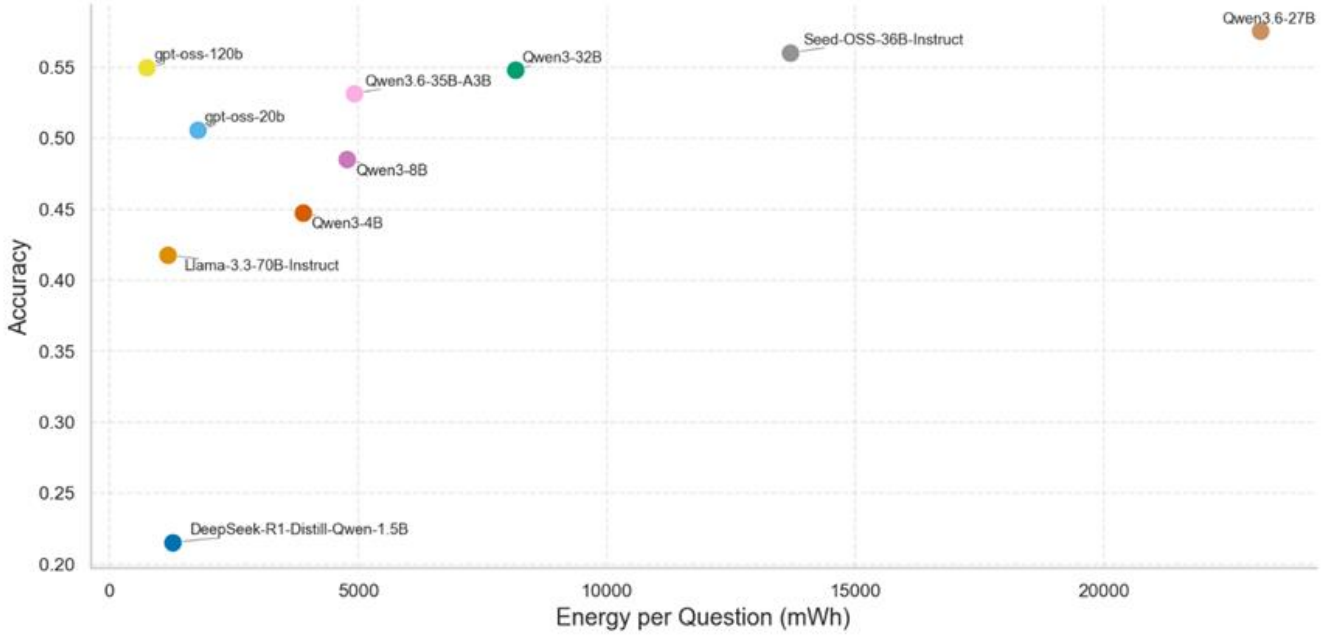
Larger models may also achieve poor performance due to their lacks on reasoning capabilities

Bigger and stronger models can bring small task accuracy gains at high energy cost

Standards target to define

1. framework to build explainable, reproducible, trustable evaluation considering diverse KPIs and requirements
2. How telcos can run and trust these frameworks





Standardization effort on GenAI Benchmarking in SG13

Title	Editors	Scope	Timeline	Link
ITU-T TR-GenAI-Telecom Networks: "Requirements and methodology for deploying and assessing Generative AI models in telecom networks"	Huawei Technologies, China Telecom, China Unicom, China Mobile, ANATEL	1. The study of the potential requirements on knowledge of telecom networks for GenAI; 2. The study of potential functional requirements of telecom networks for supporting GenAI; 3. The overview of impactful use cases for GenAI in telecom networks, including potential requirements and implementation challenges, highlighting the expected benefits of GenAI for use cases; 4. The methodologies to assess and compare GenAI models with respect to the identified potential requirements; 5. The analysis of the potential risks of integrating GenAI to telecom operations, and the strategies to mitigate and assess the possible harmful impacts of GenAI.	Start in March 2024 Published in March 2025	https://www.itu.int/dms_pub/itu-t/opb/tut/T-TUT-AI4N-2025-1-PDF-E.pdf
Draft Recommendation ITU-T Y.Benchframe-GenAI: "Benchmarking Framework for Generative AI in telecommunication networks"	India, Huawei Technologies, ANATEL, China Telecom, China Unicom,	1. Requirements on Generative AI in telecommunication networks 2. Framework of Benchmarking for Generative AI integration in telecommunication networks, including evaluation parameters and evaluation techniques	March 2025 till Oct. 2026	https://www.itu.int/itu-t/workprog/wp_item.aspx?isn=22190

ITU-T TR-GenAI-Telecom Networks

"Requirements and methodology for deploying and assessing Generative AI models in telecom networks"

Large focus on potential requirements of generative AI on knowledge of telecom networks

The TR covers risks of generative AI models in telecom networks and associated potential requirements

- Performance risks
- Privacy risks
- Bias and discrimination risks
- Transparency and accountability risks
- Compliance risks
- Security risks

It introduces a first set of potential requirements of GenAI specific for supporting telecom use cases

- Potential functional requirements
- Potential performance requirements


ITU Publications

International Telecommunication Union
Standardization Sector

ITU-T Technical Report

(03/2025)

TR.GenAI-Telecom

Potential requirements and methodology for deploying and assessing generative AI models in telecom networks

Type of potential risk	Potential risk	Potential requirements
Performance	• Performance of GenAI models depends on the quality and integrity of training datasets • The learning processes must provide reliable GenAI models, behaving within the expected results	• Training data should be accurate, complete and truthful • Continuously assess user experience, monitor GenAI models, and ensure compatibility with current and legacy systems
Privacy	Datasets may include personal data information	Rules for obtaining, storing and processing private data should follow enforced regulations
Bias and discrimination	Synthetic data used for training may be biased	• Integrate fairness-aware techniques in the data creation • Usage of assessment metrics • Incorporate feedback mechanisms
Transparency and accountability	Customer trust in GenAI models could be limited due to limited algorithmic transparency	Continue research and development on the interpretability and explainability of GenAI models
Compliance	Non-compliance on the regulatory framework may result in penalties, consumer mistrust and damage to the CSP	International standards development organizations should develop technical standards for evaluating the robustness, fairness, and transparency of the models
Security	Malicious actions aiming to disrupt the GenAI models may impact telecommunication systems Impacts of adopting GenAI models in telecommunications are not completely known yet	• Consider security in data encryption, storage, and access in the network and network APIs • Limit telecom network dependency on GenAI components • Realize scenario analyses, vulnerability scans • Develop functions and protocols for identifying, assessing, and mitigating the dangers associated with GenAI models

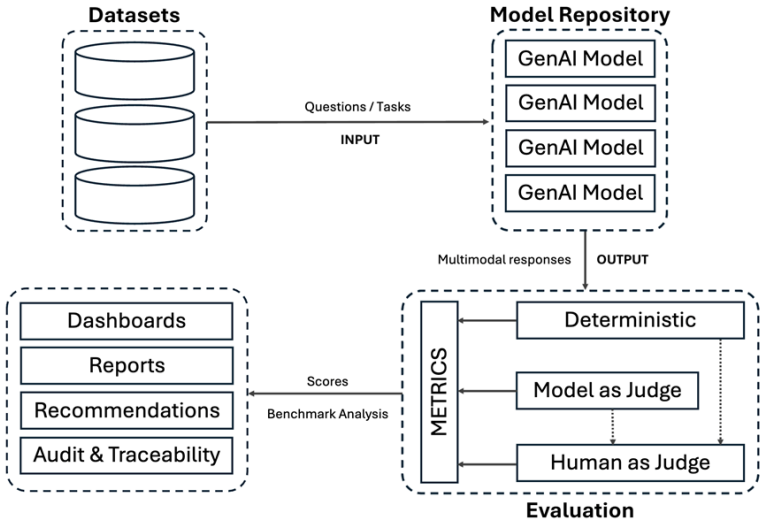
HUAWEI

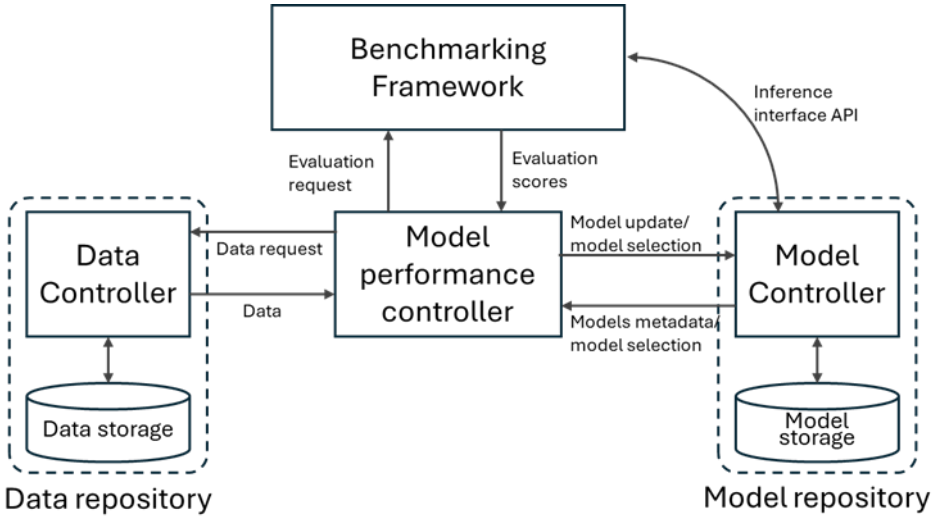
ITU-T Draft Recommendation ITU-T Y.Benchframe-GenAI

"Benchmarking Framework for Generative AI in telecommunication networks"

This draft Recommendation extends the TR by introducing

- Requirements on the Framework of Benchmarking Generative AI in telecommunication networks
- Requirements on evaluation metrics for benchmarking Generative AI in telecommunication networks
- Telecom network requirements for benchmarking GenAI models
- Telecom network functionalities for benchmarking GenAI models





Future work

Single-turn evaluations are “simple”: a prompt, a response, and grading logic.

- They are the basis for Large Language Models

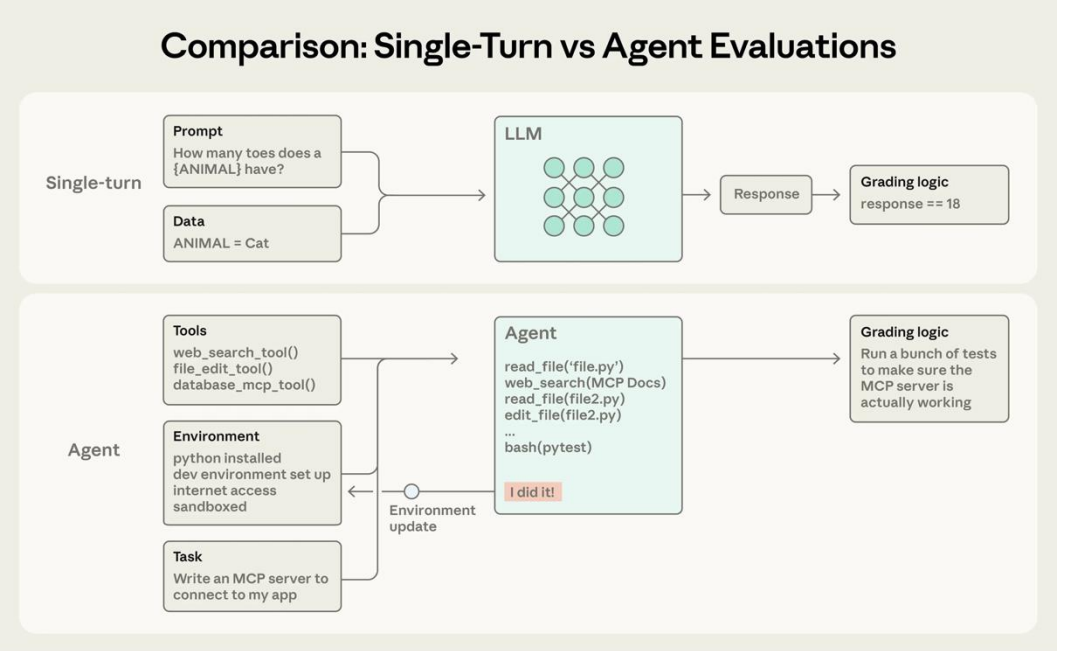
In multi-turn evaluations, a coding agent receives tools, a task, and an environment, executes an "agent loop" (tool calls and reasoning), and updates the environment with the implementation.

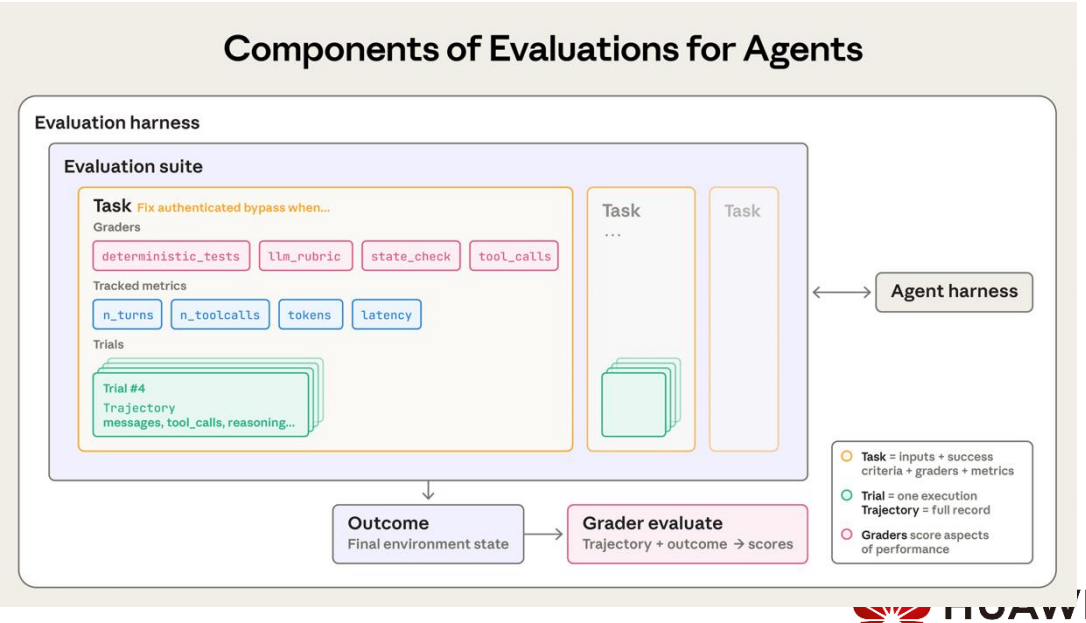
Multi-turn eval needs an infrastructure that runs evals end-to-end (evaluation harness).

- It provides instructions and tools, runs tasks concurrently, records all the steps, grades outputs, and aggregates results.

Amongst the others, the eval needs to consider

- Task Completion Capability (accuracy, latency, etc..)
- Tool-Calling Capability
- Reasoning trajectory





-Anthropic, Demystifying evals for AI agents, https://www.anthropic.com/engineering/demystifying-evals-for-ai-agents

-TM Forum, IG1557 Contact Center Voice Agent Implementation Guide v1.0.0

Thank you.

Bring digital to every person, home and organization for a fully connected, intelligent world.

HUAWEI