Technical Report

ISO/IEC TR 42106

First edition
2026-07

Information technology — Artificial intelligence (AI) — Overview of differentiated benchmarking of AI system quality characteristics

Technologies de l'information — Intelligence artificielle (IA) — Aperçu de l'analyse comparative différenciée des caractéristiques de qualité des systèmes d'IA

Reference number
ISO/IEC TR 42106:2026(en)

COPYRIGHT PROTECTED DOCUMENT

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives or www.iec.ch/members_experts/refdocs).

ISO and IEC draw attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO and IEC take no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO and IEC had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents and https://patents.iec.ch. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html. In the IEC, see www.iec.ch/understanding-standards.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 42, *Artificial intelligence*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html and www.iec.ch/national-committees.

Introduction

Artificial intelligence (AI) systems are diverse in nature and heterogeneous in terms of their potential effect on consumers and third-parties. For example, a company's use of disposable income estimates using a linear regression model to serve ads targeted at different socio-economic profiles reflects a system with low complexity and potential impact. Conversely, a bank's use of disposable income estimates using a large neural network model to make housing loan decisions reflects a system with both high complexity and high impact.

Benchmarking is often used to compare quality characteristics of software systems against reference or target values. Given the diverse nature and heterogeneous impact of AI systems, such reference or target values can differ widely across systems and across deployments of similar systems across contexts of use. To accommodate this diversity and heterogeneity, differentiated benchmarking can target different quality characteristics at different reference values. This document reviews AI management frameworks for their ability to offer guidance for differentiated benchmarking of AI system quality characteristics.

By evaluating frameworks for specifying differing levels of benchmarks and benchmarking, commensurate with expected social impact of AI systems, this document identifies gaps in current frameworks that, when filled, can yield guidance for differentiated benchmarking of AI systems, which can help rationalize the standardization implementation effort of AI providers, while maintaining system trustworthiness for AI customers and AI partners.

Information technology — Artificial intelligence (AI) — Overview of differentiated benchmarking of AI system quality characteristics

1 Scope

This document provides an overview of conceptual frameworks for graded benchmarking of artificial intelligence (AI) system quality characteristics. The aim is to examine the feasibility of using differentiated benchmarking of quality characteristics based on the complexity and context of use of an AI system.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 22989, *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*

ISO/IEC 23053, *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

ISO/IEC 25059, *Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems*

Bibliography

[1] ISO/IEC 25059, *Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems*

[2] ISO/IEC 29155-1:2017, *Systems and software engineering - Information technology project performance benchmarking framework - Part 1: Concepts and definitions*

[3] ISO 41011:2024, *Facility management — Vocabulary*

[4] ISO 14031:2021, *Environmental management — Environmental performance evaluation — Guidelines*

[5] ISO 17258:2015, *Statistical methods — Six Sigma — Basic criteria underlying benchmarking for Six Sigma in organisations*

[6] ISO 14050:2020, *Environmental management — Vocabulary*

[7] ISO 21678:2020, *Sustainability in buildings and civil engineering works — Indicators and benchmarks — Principles, requirements and guidelines*

[8] ISO 21931-1:2022, *Sustainability in buildings and civil engineering works — Framework for methods of assessment of the environmental, social and economic performance of construction works as a basis for sustainability assessment — Part 1: Buildings*

[9] ISO/IEC/IEEE 24765:2017, *Systems and software engineering - Vocabulary*

[10] ISO/TS 18667:2018, *Space systems — Capability-based Safety, Dependability, and Quality Assurance (SD&QA) programme management*

[11] ISO 20468-1:2018, *Guidelines for performance evaluation of treatment technologies for water reuse systems — Part 1: General*

[12] ISO 13053-2:2011, *Quantitative methods in process improvement — Six Sigma — Part 2: Tools and techniques*

[13] ISO/IEC 25040, *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality evaluation framework*

[14] ISO/IEC 18520:2019, *Information technology — Computer graphics, image processing and environmental data representation — Benchmarking of vision-based spatial registration and tracking methods for mixed and augmented reality (MAR)*

[15] ISO/TR 24514:2018, *Activities relating to drinking water and wastewater services — Examples of the use of performance indicators using ISO 24510, ISO 24511 and ISO 24512 and related methodologies*

[16] ISO 14644-16:2019, *Cleanrooms and associated controlled environments — Part 16: Energy efficiency in cleanrooms and separative devices*

[17] ISO 10010:2022, *Quality management — Guidance to understand, evaluate and improve organizational quality culture*

[18] ISO 10014:2021, *Quality management systems — Managing an organization for quality results — Guidance for realizing financial and economic benefits*

[19] ISO 30400:2022, *Human resource management — Vocabulary*

[20] ISO 32210:2022, *Sustainable finance — Guidance on the application of sustainability principles for organizations in the financial sector*

[21] ISO/IEC TS 25058, *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guidance for quality evaluation of artificial intelligence (AI) systems*

[22] ISO 24523:2017, *Service activities relating to drinking water supply systems and wastewater systems — Guidelines for benchmarking of water utilities*

[23] ISO 24513:2019, *Service activities relating to drinking water supply, wastewater and stormwater systems — Vocabulary*

[24] ISO/IEC 29155-1, *Systems and software engineering — Information technology project performance benchmarking framework — Part 1: Concepts and definitions*

[25] ISO/IEC 25000, *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guide to SQuaRE*

[26] ISO/IEC TS 4213, *Information technology — Artificial intelligence — Assessment of machine learning classification performance*

[27] Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer.

[28] Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems (2nd ed.). O'Reilly Media.

[29] Gulli, A., & Pal, S. (2019). Advanced deep learning with TensorFlow 2 and Keras (2nd ed.). Packt Publishing.

[30] Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and practice (3rd ed.). OTexts.

[31] Jurafsky, D., & Martin, J. H. (2025). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models (3rd ed.).

[32] Tunstall, L., von Werra, L., & Wolf, T. (2022). Natural language processing with transformers. O'Reilly Media.

[33] IEEE 2937, *IEEE Standard for Performance Benchmarking for Artificial Intelligence Server Systems*

[34] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 248–255). IEEE.

[35] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324.

[36] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3), 211–252.

[37] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In Computer Vision – ECCV 2014 (Lecture Notes in Computer Science, Vol. 8693, pp. 740–755). Springer.

[38] Phillips, P. J., Grother, P., Micheals, R., Blackburn, D. M., Tabassi, E., & Bone, M. (2003, October). Face recognition vendor test 2002. In 2003 IEEE International SOI Conference. Proceedings (Cat. No. 03CH37443) (p. 44). IEEE.

[39] Tang, F., Gao, W., Zhan, J., Lan, C., Wen, X., Wang, L., Ye, H, et al. (2021, March). AIBench training: Balanced industry-standard AI training benchmarking. In 2021 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS) (pp. 24-35). IEEE.

[40] Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Liang, P., et al. (2021, July). Wilds: A benchmark of in-the-wild distribution shifts. In International conference on machine learning (pp. 5637-5664). PMLR.

[41] Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. (2015). Generalization in adaptive data analysis and holdout reuse. Advances in Neural Information Processing Systems, 28.

[42] Leveson, N. (2020). Are you sure your software will not kill anyone?. Communications of the ACM, 63(2), 25-28.

[43] NIST AI 100-1:2023, *Artificial Intelligence Risk Management Framework*

[44] Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., Shah, N. H., et al. (2023). The shaky foundations of large language models and foundation models for electronic health records. npj Digital Medicine, 6(1), 135.

[45] Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. ACM Computing Surveys (CSUR), 55(3), 1-44.

[46] Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., Gadepally, V., et al. (2023, September). From words to watts: Benchmarking the energy costs of large language model inference. In 2023 IEEE High Performance Extreme Computing Conference (HPEC) (pp. 1-9). IEEE.

[47] H. Inan, K. Upasani, J. Chi et al., Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations, Meta Research, 2023. Available at: https://arxiv.org/pdf/2312.06674

[48] P.Liang, R. Bommasani, T. Lee et al., Holistic Evaluation of Language Models, Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford University, 2022. Available at: https://arxiv.org/pdf/2211.09110

[49] Casner, S. M., Geven, R. W., Recker, M. P., & Schooler, J. W. (2014). The retention of manual flying skills in the automated cockpit. Human factors, 56(8), 1506-1516.

[50] Casner, S. M., Hutchins, E. L., & Norman, D. (2016). The challenges of partially automated driving. Communications of the ACM, 59(5), 70-77.

[51] ISO/IEC 42001, *Information technology — Artificial intelligence — Management system*

[52] Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0).

[53] European Commission AI HLEG (2020). Assessment List for Trustworthy Artificial Intelligence (ALTAI). Available at https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342

[54] Algorithmic Impact Assessment Tool. Available at https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

[55] IEEE 7010, *IEEE Recommended Practices for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-being*

[56] Hert, C. A. (1994). Information Technology Standardization: A Classification Process?. Advances in Classification Research Online, 95-110.

[57] Shen, L. (2014). The NIST cybersecurity framework: Overview and potential impacts. Scitech Lawyer, 10(4), 16.

[58] Opinion of the German Data Ethics Commission (2019). Available at: https://www.bfdi.bund.de/SharedDocs/Downloads/EN/Datenschutz/Data-Ethics-Commission_Opinion.pdf?__blob=publicationFile&v=1.

[59] EU AI Act (2024). Available at https://artificialintelligenceact.eu/

[60] AlgorithmWatch (2021). Automated Decision-Making Systems in the Public Sector: An Impact Assessment Tool for Public Authorities. Available at https://algorithmwatch.org/en/wp-content/uploads/2021/06/ADMS-in-the-Public-Sector-Impact-Assessment-Tool-AlgorithmWatch-June-2021.pdf

[61] Platform for the Information Society (2018). Artificial Intelligence Impact Assessment Framework. Available at https://ecp.nl/wp-content/uploads/2019/01/Artificial-Intelligence-Impact-Assessment-English.pdf

[62] European Law Institute (2022). Model Rules on Impact Assessment of Algorithmic Decision-Making Systems Used by Public Administration. Available at https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ELI_Model_Rules_on_Impact_Assessment_of_ADMSs_Used_by_Public_Administration.pdf

[63] Cox, Anthony (2008), What's Wrong with Risk Matrices, Risk Analysis 28(2):497–512.

[64] Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. Psychological science, 15(8), 534-539.

[65] Yoo, S., Gregorian, D., Kopeikin, A., Leveson, N. (2023). Improving the Risk Matrix. In: Karakoc, T.H., Yilmaz, N., Dalkiran, A., Ercan, A.H. (eds) New Achievements in Unmanned Systems. ISUDEF 2021. Sustainable Aviation. Springer, Cham.

[66] ITU-T F.748.11, *Metrics and evaluation methods for a deep neural network processor benchmark*

[67] ITU-T F.748.12, *Deep learning software framework evaluation methodology*

[68] ISO/IEC TR 12182, *Systems and software engineering — Framework for categorization of IT systems and software, and guide for applying it*

[69] ISO/IEC 25002, *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model overview and usage*

[70] ISO/IEC TR 24030:2024, *Information technology - Artificial intelligence (AI) - Use cases*

[71] ISO/IEC 42001:2023, *Information technology — Artificial intelligence — Management system*

[72] ISO/IEC 22989:2022, *Information technology - Artificial intelligence - Artificial intelligence concepts and terminology*

[73] ISO/IEC TR 24028:2020, *Information technology - Artificial intelligence - Overview of trustworthiness in artificial intelligence*

[74] ISO/IEC TR 5469, *Artificial intelligence — Functional safety and AI systems*

[75] ISO/IEC TS 6254, *Information technology — Artificial intelligence — Objectives and approaches for explainability and interpretability of machine learning (ML) models and artificial intelligence (AI) systems*