AI BENCHMARKING IN BUSINESS

Aleksandar Ivanov
Burgas Free University

Abstract: *This paper presents a concise review of current AI benchmarks relevant to business applications, focusing on their design, evaluation metrics, and practical relevance. It examines how benchmarks in areas such as natural language processing, predictive analytics, and decision automation align with real-world business needs, including accuracy, scalability, fairness, and interpretability. The study highlights gaps between academic benchmarks and enterprise use cases, emphasizing the need for more context-aware and industry-specific evaluation frameworks. Ultimately, the paper aims to guide researchers and practitioners in selecting or designing benchmarks that better reflect the complex demands of AI deployment in business environments.*

Keywords: *artificial intelligence, benchmarking, business metrics*

1. Introduction

AI benchmarking is the practice of systematically evaluating and comparing the performance of Artificial Intelligence (AI) systems using standardized tasks, datasets, and metrics. It serves as a ground for measuring progress, strengths and weaknesses analysis of different models, and guiding both academic and business deployment. Robust benchmarks provide objective criteria for comparison, thus ensuring reproducibility and fairness across experiments. Benchmarks play essential role in innovation and defining the future outlook in the AI integration. As AI influences various domains, standardized benchmarking becomes increasingly important for aligning technical advancements with real-world needs and ethical considerations. [1-8]

II. Metrics and datasets

A. Technical metrics

To evaluate AI performance in different machine learning (ML) approaches various numerical metrics can quantify output accuracy and quality. In classification, accuracy, precision, recall, F1-score, and confusion matrices give insight of how well a model distinguishes classes and determines specific error types. Regression tasks use mathematical metrics like mean squared error, root mean squared error, and R-squared to assess alignment between predicted and actual values. Clustering is evaluated with silhouette score, adjusted Rand index, and mutual information, which measure similarity to ground truth or cluster cohesion. These metrics are essential for validating, comparing, and fine-tuning models, and they also can ensure they generalize well beyond training data.

Benchmark datasets are central for evaluating AI models. They must include standardized, publicly available data that support consistent testing across variety of methods. For classification, datasets like ImageNet (for image recognition) [9] and IMDb (for sentiment analysis) [10] provide large, labeled samples.

Regression tasks use benchmarks such as Boston [11] or California Housing. More specifically, for image classification, besides ImageNet, CIFAR-10/100 [12], and MNIST

[13] are common, containing large sets of labeled images. In natural language processing, GLUE (General Language Understanding Evaluation) [14], SQuAD (Stanford Question Answering Dataset) [15] benchmark models in language understanding and sentiment classification. In clustering, the Iris dataset is a classic example for evaluating unsupervised grouping, while 20 Newsgroups [16] provides high-dimensional text data suitable for clustering and topic modeling. There is a number of other datasets for clustering as well [17]. These datasets are publicly available and well-documented.

B. Fairness metrics

Fairness in AI models means providing all social groups with relevant, trustworthy information and consistent interaction, regardless of characteristics like gender, ethnicity, race, etc. Numerical and categorical measures identify biases in AI outcomes by checking disparate impact, equal opportunity, statistical parity and other indicators. Biased models can cause serious harm in areas like loan approvals, hiring algorithms, and medical diagnostics. Bias mitigation involves defining sensitive attributes (e.g., gender, race), selecting fairness metrics based on domain and use case, and measuring bias using established tools. Mitigation can occur at various data pipeline stages: pre-processing (e.g., reweighting data), in-processing (e.g., adversarial debiasing), and post-processing (e.g., equalizing outcomes).

Table 1 summarizes most popular fairness metrics.

Table 1. Fairness measures compared

Metric	Description	Goal
Statistical Parity	Equal probability of favorable outcome across groups	All groups should have similar outcomes
Equalized Odds	Equal true/false positive rates (TPR/FPR) across groups	No bias in model's success/failure rates
Demographic Parity	Outcome is independent of sensitive attribute	No correlation between outcome and demographics
Calibration	Predicted probabilities mean the same thing across groups	Predictions are equally reliable

C. Business metrics relevant to use cases

Business has adopted AI solutions in the last decade at a large scale. AI technologies are deeply integrated in a broad spectrum of fields and all aspects of the business processes, In the following paragraph a comprehensive list of AI solutions is presented, structured by means of business aspects. The list includes measures relevant to AI performance in the described context.

Customer service and support – a busssiness aspect that incorporates chatbots and virtual assistants to communicate with clients, extensively used in online services. Some of the software tools used in business are Dialogflow [19], IBM Watson Assistant [20], Rasa [21]. Some relevant metrics are accuracy and chatbot response time to user queries, user feedback measured by Customer Satisfaction – CSAT, First Response Resolution – FCR (the percentage of issues resolved within the first interaction).

Sales and marketing automation – it involves automated evaluation of user interactions and attitudes with the company services and products. It measures user engagement and

interaction intensity and it can inform and enable company policies relevant to measured indicators. Tools used in this area of AI integration are Salesforce Einstein [22], HubSpot [23], Marketo [24]. The list of indicators that can be automatically monitored via AI tools includes the rate of conversion from visitor to paying client, Lead Scoring Accuracy, Customer Lifetime Value – CLV (prediction of a customer's total value over their entire relationship with the company); Click-Through Rate – CTR (for digital ads or emails, the rate at which recipients click on AI-generated ads or messages; engagement metrics).

Fraud detection and risk management – a key aspect of businesses that provides sensitive services related to finance, user data, security and medicine. With the increasing use of online banking, online payments, digital money and cryptocurrency, transaction safety is not merely an advantage, but an imperative. Common tools that provide security services are SAS [25], FICO [26], IBM Trusteer [27], Darktrace [28]. Some of the benchmarks used to evaluate safety in AI systems are: the count of correctly flagged fraudulent transactions, Time-to-Detection of fraudulent activity, AI model explainability. Risk Prediction Accuracy, precision and recall for fraudulent transactions, False Positive Rate - FPR (the rate at which legitimate transactions or claims are incorrectly flagged as fraudulent); True Positive Rate – TPR (the percentage of fraudulent transactions accurately flagged as fraudulent).

Supply chain optimization – AI tools are capable of analysis and optimization of logistic problems. Supply chains are crucial for delivering timely and trustworthy services to customers and they are also key factors for determining end prices of products. Optimization of supply chains can reduce company expenses and generate profits. AI can predict future demands based on historical data, dynamically reflect rapid changes in demand, optimise technical parameters of delivery and storage and even propose new strategies for decision making. Corporate solutions that can optimise supply chains are Llamasoft [29], Blue Yonder [30], Oracle Supply Chain Management Cloud [31]. Common measures to optimise are precision of prediction for future product demand, often measured as Mean Absolute Percentage Error (MAPE), Order Fulfillment Time, Inventory Turnover Ratio, Supply Chain Visibility, Lead Time: the time between receiving an order and shipping it.

Human Resources and talent management – AI in HR is aimed at automating repetitive tasks such as resume screening and onboarding. Furthermore, AI-powered analytics can provide valuable insights into workforce trends, optimise talent management, and enhance the overall employee experience. Commercial tools used in HR are Workday [32], HireVue [33], Pymetrics (now Harver) [34], LinkedIn Talent Insights [35]. It is crucial for systems to be bias free. In the modern world fairness and social justice are main factors for building companies' trust as seen with recent events (boycotts of brands and institutions for discriminative policies). Measures that track effectiveness in AI assisted hiring activities are time to hire, Candidate Match Accuracy, Employee Retention Rate, Employee Engagement Scores, Training Effectiveness.

Product recommendations and personalization – recommender systems are one of the most widely adopted AI systems. They are present in nearly every online service – online stores, content sharing platforms, streaming and gaming services, etc. They are the driving force behind personalized advertising and can directly influence the cost of advertising. Many companies use the strategy of profiting from ads and providing free services to end users, so the efficiency of recommender systems is the most important tool to generate profits. Popular solutions are Amazon Personalize [36], Dynamic Yield [37], Adobe Target [38], Google Recommendations AI [39]. Customer satisfaction, generated revenue per user, recommendation accuracy are among the measures that evaluate efficiency of recommendations.

Predictive maintenance – many companies, especially in IT sector, are having large technological facilities – datacenters, corporate networks, internal infrastructure. The maintenance of such large facilities can be challenging and it can be improved by automation. Automated smart tools can predict and avoid technical failures and thus preventing losses, accidents and delivering uninterrupted user experience; it also can minimize the cost of usage by responding to changes in workloads. IBM Maximo [40], Uptake [41], Siemens InsightHub [42]. Several technical indicators can be monitored with automated systems.

- **Mean Time Between Failures (MTBF)**: the average time between two successive failures of equipment
- **Downtime reduction**: the amount of time equipment is offline due to unexpected failures
- **Maintenance cost savings**: The reduction in maintenance costs through proactive repairs and optimization, compared to reactive maintenance

Other important indicators to measure are failure prediction accuracy and repair time.

Business Intelligence (BI) and data analytics – data analysis is crucial for all medium and large scale businesses, It is related to specific tasks, planning, cost minimization, risk assessment and others. Some software tools are Tableau [42], Power BI [43], Qlik [44], Google Cloud AI [45]. Speed of data procesing and decision-making and accuracy of analysis (for example A/B tetsing) along with user adoption rate are vital indicators to evaluate AI systems.

Document processing and Natural Language Processing (NLP) – the tasks of infromation retrieval and data evaluation is increasingly getting automated with NLP tools. Such kind of automation could potentially benefit wide range of businesses, Text processing is the essential taks in opinion mining and people metrics and its automation can be transformative for the interested party. Google Cloud Natural Language API [46], IBM Watson NLP [20], Amazon Comprehend [47], and commercial LLMs are used by multitude of companies to perform NLP tasks. LLM efficiency can be measured by many benchmarks, such as acuracy and speed of text and specifically named entity recognition (NER); sentiment analysis and summarization quality can also be tested.

Financial forecasting and investment – decision making in companies is usually performed considering a timeframe that includes both past and future time periods. Prioritazing profits in time is crucial to ensure the tradeoff between long term stability and agility in rapidly changing situations. Some of the tools that provide professional financial forecasting are Bloomberg Terminal [48], AlphaSense [49], QuantConnect [50], DataRobot [51]. Here are some specific metrics used in benchmarking finacial tools.

- **Sharpe Ratio**: in algorithmic trading, the Sharpe ratio measures risk-adjusted returns.
- **Portfolio performance**: how well AI-driven investment strategies outperform traditional approaches.
- **Volatility prediction**: the accuracy of AI models in predicting market volatility.
- **Drawdown**: the maximum loss from a peak to a trough in investment portfolios.

D. General business metrics

The relationship between AI and business is bidirectional. AI facilitates the calculation of business metrics like Return-Of-Investment (ROI), Net Profit Value (NPV), Internal Return Rate (IRR), and Key Performance Indicators (KPIs) through predictive analytics, automation, and real-time insights. At the same time, these financial measures are used to

evaluate AI investments [52]. For instance, companies assess the ROI of AI-driven services by analyzing costs, projected gains, and performance indicators like accuracy and resolution time.

Return on Investment is a financial metric used to evaluate the efficiency or profitability of an investment by expressing gain or loss relative to the initial cost. It's typically calculated as net profit divided by investment cost, then multiplied by 100 to obtain a percentage. ROI is broadly used in business and finance to compare potential returns of projects, which facilitates decision-makers in prioritizing resource allocation. In AI, ROI also includes non-financial benefits such as time savings, improved accuracy, or better client experience, which may indirectly improve financial performance. A high ROI indicates significant benefits relative to cost, while a low or negative ROI signals poor or inefficient returns [52].

Total Cost of Ownership is an estimate that calculates the full cost of acquisition, operation, and maintenance of a product or system over its whole lifecycle [53]. In addition to purchase price, TCO includes direct costs like hardware, software, licensing, and implementation, as well as indirect expenses such as staff training, support, downtime, upgrades, and eventual decommissioning. For IT and AI solutions, TCO is key for understanding long-term value and for reasonable budgeting, especially when comparing on-premises infrastructure to cloud services or different vendors. Since it accounts for both visible and hidden costs, TCO offers a more realistic foundation for strategic decision-making and prevents underestimating the true investment required.

Key Performance Indicators are measurable values which organizations use to assess the effectiveness of activities in achieving strategic and operational goals [53]. They offer an objective way to assess progress and outcomes. Their use can enable informed, data-driven decisions. KPIs may be financial, such as revenue growth or profit margins, or non-financial, like customer satisfaction, employee retention, or system uptime, depending on the objective. In AI, KPIs might include model accuracy, latency, cost savings, or user adoption. Well-defined KPIs track accountability, and drive improvement by making success criteria explicit and actionable.

III. State-of-the-art benchmark solutions

A. Fairness benchmarks and mitigation algorithms

In this segment there is a brief overview of popular techniques to detect and mitigate biases as well as standardized benchmarks. Here is a short list of popular algorithms:

Reweighing (Kamiran and Calders, 2012) – a popular preprocessing technique that adjusts the weights of training samples to balance representation across protected groups prior to avoid bias [18].

Equalized odds postprocessing (Hardt et al., 2016) – a postprocessing method that enforces equal false positive and false negative rates across groups, serving as a benchmark for fairness in classification. It is applied after training [54].

Adversarial debiasing (Zhang et al., 2018) is an in-processing method using adversarial learning to remove sensitive attribute information while preserving predictive performance. It involves two neural networks with opposing goals: one classifier predicts data labels, and an adversary network tries to infer the sensitive attribute from the classifier's output [55].

Exponentiated gradient reduction (Agarwal et al., 2018) – a reduction approach that frames fair classification as a constrained optimization problem, balancing accuracy and fairness [56].

Older methods like Reject Option Classification and Prejudice Remover Regularizer (2012) laid early groundwork in post- and in-processing fairness. Later approaches like *Optimised preprocessing* and *Calibrated equalized odds* refined foundational ideas for improved calibration and optimization. Recent innovations such as *Fair data adaptation* and *Sensitive subspace robustness* explore fairness through latent spaces and invariant representations. Advanced techniques – including *Rich subgroup fairness*, *Meta-algorithms, grid search reduction*, and *Differential fairness* – reflect a shift toward more granular fairness definitions and adaptive mitigation strategies. These are paired with fairness metrics like selection and error rate – based group fairness, sample distortion measures, and indices such as the *Generalized entropy index*, *Bias amplification*, and *Bias scan*, enabling evaluation beyond binary fairness notions [57].

IBM AI Fairness 360 (AIF360) is an open-source toolkit developed to detect, understand, and mitigate bias in ML models. It offers 70+ fairness metrics across different stages of the AI lifecycle and includes algorithms to reduce unfair outcomes during data preprocessing, model training, and prediction. AIF360 supports various fairness definitions and lets users explore trade-offs between accuracy and fairness. It was designed for transparency and accountability. It features rich documentation, tutorials, and compatibility with standard ML frameworks like TensorFlow, Keras, Scikit-learn, and PyTorch. It is licensed under Apache 2.0. All of the metrics mentioned above are used in AIF360 [58][59].

Aequitas is an open-source fairness audit toolkit developed by the University of Chicago’s Center for Data Science and Public Policy to help data scientists, policymakers, and stakeholders assess whether ML models produce equitable outcomes across social groups. It focuses on public policy and social impact applications. The metrics offered are interpretable and highlight disparities across subgroups defined by race, gender, or age. Aequitas provides a user-friendly interface and detailed reports to help identify biases and understand their implications without requiring deep technical expertise. It includes a Python toolkit that integrates easily with data science frameworks. The benchmark is distributed as open-source under the MIT license [60].

Fairlearn is an open-source, community-driven project for bias detection. It includes several benchmark datasets such as the ACS Income dataset, Boston Housing dataset, UCI Bank Marketing dataset, Credit Card dataset, and Diabetes 130-Hospitals dataset – suitable for testing both regression and classification models. Fairlearn supports tests for over 10 bias metrics and provides algorithms for reductions that mitigate disparity by casting disparity constraints as Lagrange multipliers. This results in reweighting and relabeling input data, reducing the problem to standard ML training [61].

ETH Zurich researchers have developed a wide range of AI benchmarks and tools across multiple domains. They created a compliance checker to evaluate how well AI models, including LLMs, meet the requirements of the EU AI Act [62]. For AI safety it is tested whether a LLM agents can generate adversarial attacks that bypass known defenses. For engineering, they benchmarked sparse Polynomial Chaos Expansions (PCE) [63] to assess their applicability to various problems. In robotics and computer vision, the ETH3D SLAM benchmark [64] supports evaluation of mono, stereo, and RGB-D SLAM algorithms using video data with ground truth. The group also developed the Education-Employment Linkage Index (EELI) [65] to quantify how vocational education systems align with employment needs. In software engineering, ETH is building a benchmark to test LLMs’ ability to infer program specifications. To support hardware innovation, they introduced PrIM [66], a benchmark suite to measure Processing-In-Memory (PIM) architecture performance under different workloads [67][68].

The abovementioned European Commission's AI Ethics Guidelines [62] offer a framework focused on transparency, accountability, fairness, and data protection, emphasizing explainability, bias prevention, and GDPR compliance, especially important for high-risk sectors like finance and healthcare. The OpenAI Charter highlights safety, ethical design, and alignment with human values, addressing long-term societal impact, risks, and governance in AI development.

B. Technical and business benchmarks

MLPerf is a widely adopted benchmark suite designed to measure the performance of ML hardware, software, and services. It provides standardized, fair, and reproducible benchmarks across a broad set of ML tasks and workloads, helping users compare how efficiently different systems train or run ML models. MLPerf covers key areas: training (how fast a system can train a model from scratch), inference (how quickly it can run predictions with a trained model), and use cases like image classification, object detection, natural language processing, and recommendation systems. Maintained by a consortium of industry leaders, academia, and research institutions–including Google, NVIDIA, Intel, AMD, and others, MLPerf is regularly updated to reflect advances in ML models and hardware architectures [69-74].

TPCx-AI is a standardized benchmark by the Transaction Processing Performance Council (TPC) for evaluating AI system performance across end-to-end ML pipelines. Unlike benchmarks focused only on inference or training, TPCx-AI spans the full lifecycle–from data ingestion and preprocessing to training, evaluation, and deployment. It offers fair, repeatable, and verifiable assessments of how systems (hardware, software, infrastructure) deal with real-world AI workloads. It is built around 40 diverse ML problems—such as fraud detection, loan prediction, image classification, time series forecasting, and text classification. It uses standard tools like XGBoost, LightGBM, and scikit-learn. TPCx-AI evaluates the complete pipeline, including ETL, feature engineering, and deployment simulation, using metrics like accuracy and F1-score. Its main metric, AIUC (AI Useful Completions per second), reflects throughput, latency, and cost-performance. The benchmark is platform-agnostic and runs on cloud platforms, on-prem GPUs, CPUs, or hybrid systems [75].

MasakhaNER is a novel dataset designed for NER tasks across multiple African languages. It includes annotations for entities such as people, organizations, locations, and dates. The dataset supports over 20 African languages, including Amharic, Nigerian Pidgin, Swahili, Yoruba, among others. The MasakhaNER 2.0 update introduced Africa-centric transfer learning approaches, which can improve performance by selecting optimal source languages for transfer learning. AfriSenti is the largest sentiment analysis dataset for underrepresented African languages, containing over 110,000 annotated tweets in 14 languages. It was utilized in the SemEval 2023 Task 12, focusing on sentiment analysis for African languages [76].

RobustBench is another standardized benchmark for evaluating adversarial robustness in ML models, particularly for image classification. It counters inflated robustness claims by providing a consistent, reliable evaluation framework using AutoAttack—a suite of white-box and black-box attacks – for rigorous assessments. RobustBench includes leaderboards comparing model performance under different threat models (e.g., $\ell\infty$, $\ell 2$ perturbations), enabling transparent comparisons. It also maintains an open-source Model Zoo with over 80 robust models, helping researchers and practitioners aiming to test, benchmark, or extend robust architectures [77].

HELM benchmark (by Stanford CRFM) holistically evaluates LLMs across tasks, metrics, and ethical concerns. Its components – HELM Instruct (instruction-following),

AIR-Bench (safety/bias), CLEVA (multilingual), and ThaiExam (Thai language) – enable standardized comparisons of functionality, fairness, and cross-linguistic competence [78].

BIG-Bench (by Google Research) tests LLMs on 204 tasks (linguistics, math, bias, etc.) via contributions from 450+ researchers. It uses human-expert baselines to assess models from millions to billions of parameters (e.g., GPT, Google models), revealing scaling trends [79].

Hugging Face's Open LLM Leaderboard offers transparent evaluations of open LLMs using benchmarks like MMLU-Pro and TruthfulQA, testing reasoning and NLP metrics. [80].

Business benchmarks include Deloitte's AI Value Trees and McKinsey's framework, both aligning AI initiatives with strategic goals to measure outcomes.

This paper is not aimed at advertising any of the software mentioned nor business solutions. The provided links for companies sites may not be supported in the long term.

IV. Challenges and considerations

Data sensitivity and security are critical in AI benchmarking, as datasets often contain personal or proprietary information. Ensuring privacy through encryption, access controls, and anonymization align with regulations like GDPR and fosters trust by preventing distorted results. Security is vital for ethical and trustworthy AI deployment.

A major challenge is AI's rapid evolution – new models, tasks, and architectures (e.g., multi-modal reasoning) quickly outdate benchmarks. Many are narrowly focused and fail to generalize about complex real-world use-cases. Licensing and data availability also limit diverse, open benchmarks. Overfitting is also common: models may excel on specific benchmarks but lack broader utility. Hardware differences and optimization stacks further hinder fair comparisons.

Efforts continue to update benchmarks, but maintaining fairness and relevance remains difficult. Qualitative factors like user satisfaction and brand trust are becoming increasingly important. Traditional metrics often overlook ethical impacts and user experience, both key to long-term value. New benchmarking frameworks address these challenges by exploring human-centered metrics like user studies, feedback loops, sentiment analysis, and net promoter scores. Broader effects – trust, social responsibility, regulatory fit—are measured via proxies like complaint rates or audit scores. Though harder to standardize, these make evaluations more holistic, recognizing that AI success relies on trust and societal alignment.

One improvement could be interdisciplinary groups creating official standards for metrics, datasets, and algorithms. In the light of AI's rapid pace, automated or semi-automated pipelines could help benchmarks evolve with technology. Meta-learning platforms could enable AI to evaluate other models and redefine benchmarks. Advances in eXplainable AI (XAI) and transfer learning could also offer clearer, more robust frameworks for understanding and reuse.

V. Conclusions

There is a great variety of benchmarks for AI testing, but currently there is no universal and widely accepted standard neither in academic nor in the business worlds. This leads to fragmentation and non-reliable testing, which in turn can cause inefficient resource planning and profit forecasting in companies. This study clearly indicates the gaps in AI benchmarking, and it highlights the need for strategy creation and adoption.

References:

1. Thiyagalingam, J., von Laszewski, G., Yin, J., Emani, M., Papay, et al.,. AI benchmarking for science: Efforts from the MLCommons science working group. *InInternational Conference on High Performance Computing 2022* (pp. 47-64).
2. Enholm, I., Papagiannidis, E., Mikalef. P., Krogstie, J.. Artificial intelligence and business value: A literature review. *Information Systems Frontiers.* 2022
3. Thiyagalingam, J., Shankar, M., Fox, G., Hey, T. Scientific machine learning benchmarks. *Nature Reviews Physics.* 2022
4. Cowls, J., Tsamados, A., Taddeo, M., Floridi, L. A definition, benchmark and database of AI for social good initiatives. *Nature Machine Intelligence*. 2021
5. Liang, W., Tadesse, G., Ho, D., Fei-Fei L, Zaharia, M., Zhang, C., Zou, J.. Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence.* 2022
6. Chen, J., Tang, J., Qin, J., Liang, X., Li,u L, Xing, E., Lin, L. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint* arXiv:2105.14517. 2021
7. Tang, F., Gao, W., Zhan, J., Lan, C., Wen, X., Wang, L., Luo, C., Cao, Z., et al., AIBench training: Balanced industry-standard AI training benchmarking. *2021 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)* 2021
8. Hegghammer, T.. OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment. *Journal of Computational Social Science.* 2022
9. https://www.image-net.org/
10. https://developer.imdb.com/non-commercial-datasets/
11. https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset
12. https://www.cs.toronto.edu/~kriz/cifar.html
13. https://docs.ultralytics.com/datasets/classify/mnist/
14. https://gluebenchmark.com/
15. https://rajpurkar.github.io/SQuAD-explorer/
16. https://archive.ics.uci.edu/dataset/113/twenty+newsgroups
17. https://clustering-benchmarks.gagolewski.com/index.html
18. Kamiran, F., Calders, T., Pechenizkiy, M., Techniques for discrimination-free predictive models. *Discrimination and privacy in the information society: data mining and profiling in large databases.* 2013
19. https://dialogflow.cloud.google.com/#/getStarted
20. https://www.ibm.com/watson
21. https://rasa.com/
22. https://www.salesforce.com/eu/artificial-intelligence/
23. https://www.hubspot.com/
24. https://business.adobe.com/bg/products/marketo.html
25. https://www.sas.com/en_us/solutions/ai.html
26. https://www.fico.com/en
27. https://www.ibm.com/trusteer
28. https://www.darktrace.com/
29. https://www.coupa.com/products/supply-chain-design/
30. https://blueyonder.com/
31. https://www.oracle.com/scm/
32. https://www.workday.com/en-us/artificial-intelligence.html
33. https://www.hirevue.com/

34. https://harver.com/
35. https://business.linkedin.com/talent-solutions/talent-insights
36. https://aws.amazon.com/personalize/
37. https://www.dynamicyield.com/ai/
38. https://business.adobe.com/products/target.html
39. https://cloud.google.com/use-cases/recommendations
40. https://www.ibm.com/products/maximo
41. https://uptake.com/
42. https://plm.sw.siemens.com/en-US/insights-hub/resources/faq/
43. https://www.tableau.com/
44. https://powerbi.pl/en/blog/microsoft-power-bi-en/power-bi-and-artificial-intelligence-how-ai-is-revolutionizing-data-analytics
45. https://www.qlik.com/us
46. https://cloud.google.com/natural-language
47. https://aws.amazon.com/comprehend/
48. https://www.bloomberg.com/professional/products/bloomberg-terminal/
49. https://www.alpha-sense.com/
50. https://www.quantconnect.com/
51. https://www.datarobot.com/
52. Buriev, S., KEY STATISTICAL INDICATORS IN THE IMPLEMENTATION OF INVESTMENT PROJECTS. *International Conference on Multidisciplinary Sciences and Educational Practices* 2024
53. Jansen, J.. True Costing in Logistics & Supply Chain Management: How do we make decisions based on True Economic Trade-Offs (T-ETOs)?. *London Journal of Social Sciences.* 2024
54. Hardt, M., Price, E., Srebro, N.. Equality of opportunity in supervised learning. *Advances in neural information processing systems.* 2016
55. Zhang, B., Lemoine, B, Mitchell. M.. Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* 2018
56. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H.. A reductions approach to fair classification. *InInternational conference on machine learning* 2018
57. Tan, S., Caruana, R., Hooker, G., Lou, Y.. Distill-and-compare: Auditing black-box models using transparent model distillation. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018
58. Bellamy, R., Dey, K., Hind, M., Hoffman, S., Houde, S., Kannan, K., et al. *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*.arXiv:1810.01943, 2018
59. https://github.com/Trusted-AI/AIF360
60. https://www.aequitas-project.eu/
61. https://fairlearn.org/
62. https://artificialintelligenceact.eu/
63. https://ethz.ch/content/dam/ethz/special-interest/baug/ibk/risk-safety-and-uncertainty-dam/news/Documents/PCE_Sudret(ncmdao).pdf
64. https://www.eth3d.net/slam_overview
65.https://cemets.ethz.ch/cemets-news/2016/08/education-employment-linkage-index-feasibility-study.html
65. https://github.com/CMU-SAFARI/prim-benchmarks/blob/main/run_weak.py

66. https://ces.ethz.ch/research/benchmark-instruments.html
67. https://ethz.ch/en/news-and-events/eth-news/news/2024/10/how-law-abiding-is-ai-eth-researchers-put-it-to-the-test.html
68. https://mlcommons.org/benchmarks/training/
69. Banbury, C., Reddi, V., Torelli, .P, Holleman, J., et al. Mlperf tiny benchmark. *arXiv preprint arXiv:2106.07597.* 2021
70. Reddi, V., Cheng, C., Kanter, D., Mattson, P., Schmuelling, G., Wu, C.. The vision behind mlperf: Understanding ai inference performance. *IEEE Micro.* 2021
71. Janapa Reddi, V., Kanter, D., Mattson, et al., MLPerf mobile inference benchmark: An industry-standard open-source machine learning benchmark for on-device AI. *Proceedings of Machine Learning and Systems.* 2022
72. Liu Olesiuk, Y., Hodak, M., Ellison, D., Dholakia, A.. More the merrier: comparative evaluation of TPCx-AI and MLPerf benchmarks for AI. *Technology Conference on Performance Evaluation and Benchmarking.* 2022
73. Hodak, M., Ellison, D., Dholakia, A.. Everyone is a winner: interpreting MLPerf inference benchmark results. *Technology Conference on Performance Evaluation and Benchmarking.* 2021
74. https://www.tpc.org/tpcx-ai/default5.asp
75. https://huggingface.co/datasets/masakhane/masakhaner
76. https://robustbench.github.io/
77. https://crfm.stanford.edu/helm/
78. https://paperswithcode.com/dataset/big-bench
79. https://huggingface.co/open-llm-leaderboard