Summary of "Exploring AI Benchmarks and Leaderboards"

Benchmarks in artificial intelligence are essential tools that enable standardized evaluation of model performance. A benchmark typically consists of a dataset, a set of tasks, evaluation metrics, and reference models. It allows for fair and reproducible comparison between models, fostering innovation and progress in field.

The main goals of benchmarks include:

- Standardizing evaluations (Ensuring all models are tested under the same conditions),
- Promoting scientific reproducibility (Enabling other researcher to replicate results),
- Measuring technological advancements (Tracking progress over time in model performance),
- Encouraging research through globally shared challenge (Fostering collaboration and competition within the AI community).

A benchmark is usually based on four fundamental components:

- Dataset: a corpus of data (images, texts, sounds, videos) on which the model will be tested.
- Task: for example, classifying an image, answering a question, or translating text.
- Metrics: these quantify performance (accuracy, F1-score, BLEU score, latency).
- Baselines: reference performances, human or algorithmic, used for comparison.

Benchmarks are categorized according to what they aim to measure. Some evaluate raw performance, others focus on robustness to noise or adversarial attacks. Some target computational or energy efficiency (MLPerf Power), while others address ethical and societal issues, such as fairness across demographic groups or bias mitigation (e.g., FairFace, Gender Shades).

Each subfield of artificial intelligence has its own iconic benchmarks. In natural language processing (NLP), for instance:

- GLUE, SuperGLUE, SQuAD, and XTREME test skills like comprehension, translation, or named entity recognition.

In computer vision:

- ImageNet, COCO, and CIFAR are well-established references.

In speech recognition and audio processing:

- LibriSpeech and VoxCeleb are frequently used datasets.

In reinforcement learning:

- OpenAI Gym and MuJoCo provide simulated environments to train and evaluate intelligent agents.

More complex approaches like **multimodality** (simultaneous processing of text, images, audio) rely on benchmark such as:

- VQA, CLIP, and MS COCO.

Other benchmark evaluate capabilities such as:

- generalization: WILDS, DomainNet,
- reasoning: BIG-Bench, ARC, MMLU,
- creativity, explainability, or alignment with human intent.

Some benchmarks also test human-AI interaction through:

- graphical user interfaces (MiniWoB++, RoboDesk),
- web navigation (WebGPT),
- automated task execution (RPA: Robotic Process Automation).

These benchmark are often accompanied by **leaderboards** public rankings of top performing models based on their scores. These rankings include:

- the name of the team,
- the methodology used,

- the metrics achieved,
- and sometimes a link to the code or scientific paper.

Leaderboards play a central role in AI research: they track progress, indicate the state of the art for a given task, and promote transparency and healthy competition among researchers. They are hosted on platforms such as:

- Hugging Face,
- Papers With Code,
- or directly on benchmark websites (GLUE, ImageNet, BIG-Bench, MMLU).

In conclusion, benchmarks and leaderboards are foundational pillars of AI research. They enable rigorous evaluation, promote innovation, ensure reproducibility, and help guarantee that developed models meet high standards in terms of performance, robustness, fairness, and practical utility.