AI Evaluation and ISO Standards: Is the Gap Closed?*

Andrea Trenta[1]

[1] UNINFO UNI TC 533 Technical Committee Artificial Intelligence, Turin, Italy

Abstract

This paper is a part of a set of papers showing how newly defined data and software quality measures can be described in ISO 25000 format. In the first group of papers [1], [2],[3], [4], we discussed with the help of some examples, the general approach of conformance when new quality measures are defined, and in the last papers [6], [7] how to build practical ISO/IEC 25000 compliant product quality measures for AI, starting from measures developed in several public projects. In this paper we analyze the feasibility of evaluation of an AI product, according AI ISO/IEC standards, through examples from existing practices. Moreover, this paper can be considered for the works in AI standardization area.

Keywords

product quality, measures, accuracy, metric, machine learning, artificial intelligence

1. Introduction

Policy makers, industries, and academia are facing the problem of building trust in AI, and trustworthyness, in turn, requires [15] that the AI product is measured, evaluated, and finally assessed. It is to be noted that an organization can accomplish measurements and evaluation based on existing practices as it is now advised and supported by [13]. Starting from this point, in the present paper we show, through a sort of reverse engineering, why measurements and evaluation of existing practices are coherent to SQuaRE and AI ISO standards and possible remaining issues.

Figure 1 gives an overview of entities that contribute to AI product assessment.

2. AI standardization (2024 update)

The issue of drafting standards for AI trustworthiness was mainly, but not only, assigned to the international

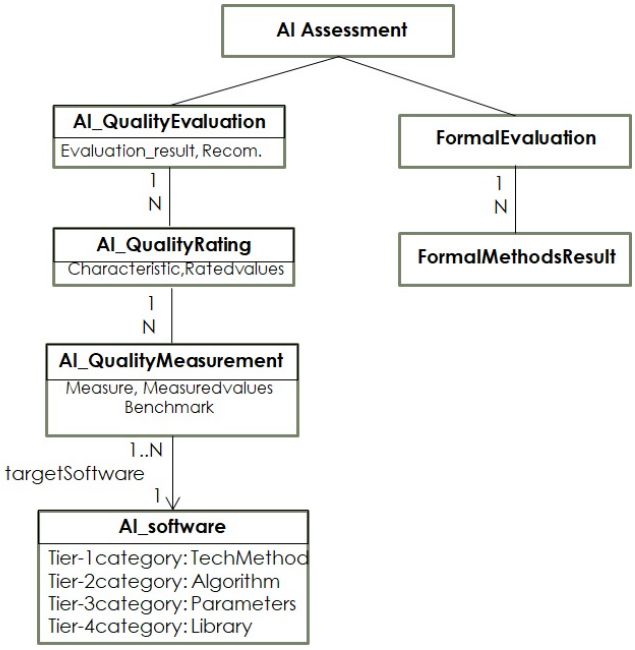

Figure 1: AI evaluation overview (UML-like)

IWESQ'24: International Workshop on Experience with SQuaRE Family and its Future Direction, December 03, 2024, Chongqing, CN
andrea.trenta@dataqualitylab.it (A. Trenta)

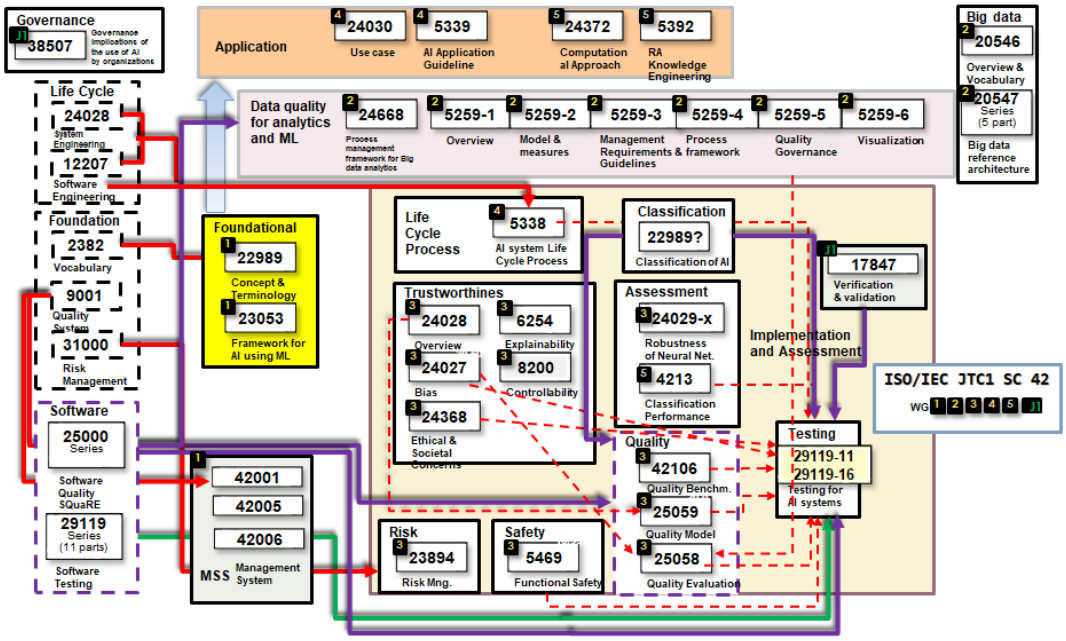

Figure 2: Non-official ISO standards by topic

standardization body ISO/IEC SC42 and to the European standardization body CEN/CENELEC JTC21 that have in charge the drafting of technical standards in support of industry and of lawful rules.

SC42 achieved important results in drafting AI standards, and this success leveraged, among the others, the foundational SQuaRE and testing standards from SC7, the implementation standards, the assessment and the management system standards. So, the work of ISO\IEC SC42 has given birth to a set of standards on AI that are covering topics such as definitions [8], software and data quality [19], testing, risk management, assessment, management system, application, according to the non-official scheme of figure 2.

It is to be noted that SQuaRE product quality standards play a central role in AI measurements and evaluation and SC42 has developed extensions [12], [14] to standards of the series ISO/IEC 25000. Indeed, the ISO/IEC 25000 series itself foresees the possibility to extend the model to specific technologies like AI, through the definition of new characteristics and new measures. This role and its reasons are also well explained in the ISO/IEC news given in https://www.iec.ch/blog/new-international-standard-ensuring-quality-ai-systems.

In this context, the assessment of product quality, possibly together with the assessment of process quality [13], will be performed on voluntary or, in the near future, on mandatory basis, in the former case to promote trustworthiness in AI systems, in the latter case to get compliance to rules [14].

3. ISO standards for AI evaluation

We focus on the topic of AI product quality evaluation and in the following, we analyze[2] a set of ISO standards and their mutual relationships relevant for organizations that develop, deploy and use AI, that are:

- ISO/IEC 42001 'AI management system' [13]
- ISO/IEC CD TS 17847 'Verification and validation analysis of AI systems' [15]
- ISO/IEC TS 25058 'Guidance for quality evaluation of artificial intelligence (AI) systems' [14]
- ISO/IEC 25059 'Quality model for AI systems' [12]
- ISO/IEC 25040 'Quality evaluation framework' [16]
- ISO/IEC AWI TR 42106[3] 'Overview of differentiated benchmarking of AI system quality characteristics' [17]

Firstly, we recall that [13], gives guidance for the management system mainly for AI product and services, as existing management systems for processes are applicable also to AI. The full picture in [13] is completed

[2] Sources are public pages of: (1) ISO Online browsing platform, available at https://www.iso.org/obp (2) ISO Online Standards directory, available at https://www.iso.org/standards.html.

[3] AWI and CD means standards 'under development'

thanks to the reference to [12], as even AI performance can be addressed through an AI quality model.

According to [15], Validation & Verification analysis for AI systems[4] is made by: (1) quality evaluation, (2) formal methods and (3) simulation. The reason why [15] was introduced in addition to the existing V&V standards, is that testing and quality evaluation are considered not exhaustive when also formal methods[5] and simulation are feasible (e.g. in case of neural networks), so that they should also be applied jointly to traditional quality evaluation for AI systems. By the way, at the moment [13] doesn't cite [15] but it could be coherently cited, when necessary, in future revisions of [13].

In all the AI standards it is recognized that the reference for AI quality evaluation is [14] that in turn is based on quality models defined in [9].

It is to be noted that for the AI set of standards, many useful concepts can be imported from [16]: (1) quality rating module, (2) recommended scoring method, (3) categorization of software, (4) quality evaluation, (5) evaluation result;

(1) Quality rating module is defined as 'set of quality measures, operational environment, and methods for conducting quality measurements and quality ratings on a specific category of target entities' and takes into account the variability of the environment and measures through a rich documentation; template rating contains a parametric description of rating method and measurement environment in the perspective they will be reused. It includes rating level setting and range of acceptance.
(2) Overall score can be a recommended scoring method and consists in a weighted linear combination of the individual scores reached from measurements of some selected measures of a group of (sub)characteristics. In this perspective, an overall quality score Qs could be[6] a sum of j-measurements Mij for each of the Wi weighted i-characteristics selected for the evaluation, and should be comparable with the relevant benchmarks[7] B_{ij}:

$$Qs = \sum_{i=1}^{n} W_i \cdot \sum_{j=1}^{m} \frac{M_{ij}}{B_{ij}}$$

(3) Categorization of software is defined as 'specific way to allocate a target system into a category'; [22] contained also the concept of differentiated levels of quality associated to an application software; for example, there are different levels of reliability required for a banking software (lower) than for a meteorological satellite (higher).
(4) Quality evaluation is composed by the following steps:
    - measurement
    - rating (single dimension)
    - rating (multiple dimensions[8])
(5) Evaluation result output can be of two types:
    - a pass/fail result or
    - a quality score

A quality score output is advised in industry context and for a user's perspective, a pass/fail output is sufficient for compliance with technical requirements or other kind of prescriptions, including certification.

The need for concepts (1), (2) and (3) is very clear and unavoidable, as AI software is highly differentiated, in the sense that it there are a lot of algorithms for each task (recognition, classification, prediction, etc.) and in turn each algorithm can behave differently due to different model, training, hyperparameters, etc.

As represented in figure 1 and detailed in [6], [7], the manifold of measures implies a deep categorization of the software under measurement, that includes even the specification of the code. This approach it is also coherent with the fact that code libraries often include their own coded measures[9] .

This need has coherently been taken into account for example in [17] where benchmarking is 'differentiated' due to the difficulty to apply an 'absolute' benchmarking.

It is to be pointed out that the issue of comparing measurements is relevant for a nonlocal quality evaluation process (i.e. products comparison) and not for measurement process: in fact, it is always possible to carry out a measurement with any metric10[10] ; on the opposite, it is challenging the comparison and replicability of the measured values, due to the extreme variability of AI products, even when performing the same task: e.g. an image classification task can be performed through thousands combinations of models, algorithms, training data, etc.

[4] It is intended an AI system that encompasses AI system components and the interaction of non-AI components with the AI system components
[5] The measurement source for formal methods is a model and not a delivered machine, on the other end, the measurement source for AI quality rating should be a deliverable source code or delivered machine and not a model.
[6] The formula is taken from [6].
[7] In the formula, B_{ij} is defined as the best value of M_{ij} (e.g., the best value of an historical series of M_{ij} measurements) and it is a normalizing factor.
[8] The multidimensional rating is better known as quality analysis.
[9] One example is Scikit-Learn https://scikitlearn.org/1.5/modules/model_evaluation.html
[10] See [1][6][18][13] for the definition of new measures.

Another widely recognized limitation to benchmarking [21][17], is when an ML implemented with neural networks uses continuous learning, its hyperparameters are evolving, and the measurement of characteristics of the NN can be different (and assessed worse or better) from the measurement taken in the initial state. This is also the reason why some AI devices are deployed and sold as "frozen" giving a guarantee to the user-buyer that the behaviour and performance value of the ML will be the same all the time.

Anyway, additional requirements (e.g. operational performance not worse than tested ones) and measurement can be satisfied, so enlarging the field of evaluation, both along the time and the post-training data and perform a further assessment of the ML in the operational mode.

4. Evaluation example: Rec. ITU-T F.748.11

The Rec. F.748.11 [24], proposes, metrics, benchmark and evaluation method for AI processors. Benchmarks are calculated both for the testing environment and for the production environment, being the latter relevant for quality evaluation.

Despite the different focus, (a processor instead of a pure software application), the approach of ITUT F.748.11 is the same of other examples as it is based on:

a) Taxonomizing of Deep Neural Networks[11] by 12 models[12]
b) Selection of 6 characteristics[13] and 11 measures
c) Definition of a quality rating module that includes 8 standard scenarios[14] ; benchmarks calculation; measurements rating.
d) Evaluation by each model and by each characteristic over single scenario

It should be noted that [24] is the only standard that details both the training scenario (e.g. ResNet_50, ImageNet) and the relevant target value, i.e. benchmark, (e.g. 74.9% Top-1 Accuracy) and this approach seems advised by [15].

So, we can conclude that both the measures and the evaluation method of ITU-T F.748.11 [24] can be accounted as compliant to ISO AI quality standards.

5. Evaluation example: Holistic Evaluation for Language Models (HELM) [23]

In [7] we showed how the measures performed in the research HELM [23] is, even if unintentionally, compliant to ISO AI quality standards.

In this paper we show that not only measures but also evaluation method of HELM is compliant to ISO AI quality standards. This conclusion is supported by the fact that HELM uses the same concepts of ISO standards, in particular, mapping entities in https://crfm.stanford.edu/helm/v0.2.2/ with entities of figure 1:

a) Taxonomizing of LLM[15] by 36 models[16]
b) Selection of 13 characteristics[17] and 57 measures
c) Definition of a quality rating module that includes 42 standard scenarios[18], then measurements rating
d) Evaluation by each model and by each characteristic over multiple scenarios, as shown in fig.3

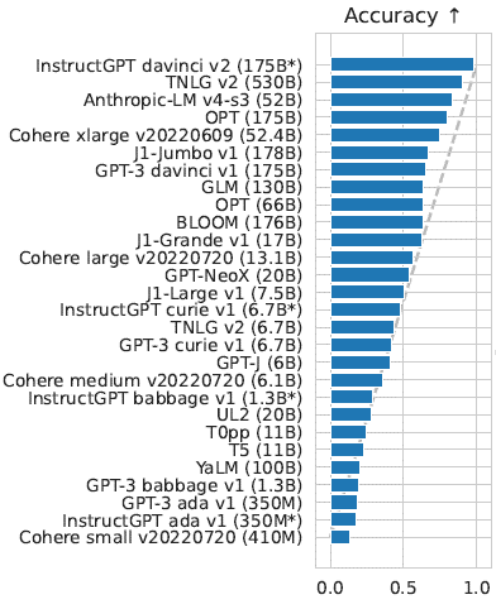

Figure 2: Accuracy over multiple scenarios (win rate) [20]

[11] Deep NN is a Techmethod in fig.1.
[12] Model is an algorithm in fig.1.
[13] Most of scenarios refer to accuracy characteristic and its measures.
[14] Standard scenarios rely on standard dataset, both in training and input prompts. A set of scenarios is in general suitable for a set of measures (e.g. CivilComments and RealToxicityPrompts dataset are suitable for toxicity measurement).

[15] LLM is a Techmethod in fig.1.
[16] Model is an algorithm in fig.1.
[17] Characteristics that are not present (e.g. toxicity) in models [5], [6], can be still handled as ISO 25000 conforming mechanism [27].
[18] Standard scenarios rely on standard dataset, both in training and input prompts. A set of scenarios is in general suitable for a set of measures (e.g. CivilComments and RealToxicityPrompts dataset are suitable for toxicity measurement).

We recall that in [7], we considered the measure of detection of toxic text[19] 19 and defined the table 1 below.

In conclusion, both the measures [20] and the evaluation method of HELM [23] can be accounted as compliant to ISO AI quality standards.

Table 1 Toxicity measure (ISO/IEC 25000 format)

ID	Txtclass-ML-1-1
Name	Text classification by toxicity
Description	Detection of toxic text in LLM input prompts
Measurement function	X= L(I, O) L= Perspective API I_1= implementation (NOTE1) O= RealToxicityPrompts (NOTE2)
NOTE1 I_1 = I_1 (method, algorithm (library, parameters), training (dataset, process)) where: method = {Large Language Models} algorithm$_1$ ={Generative Pre-trained Transformer} library = {GPT-3 davinci v1} parameters$_1$ = {parameters_set} training = {CivilComments, one-step training}	
NOTE2 https://ai2-public-datasets.s3.amazonaws.com/realtoxicityprompts/realtoxicityprompts-data.tar.gz.	

6. Evaluation example: Papers with Code [26]

Papers with Code is a resource project hosted by Meta AI Research and based on an increasing number of research papers that are mostly uploaded with code; this huge repository contains also the results of measurements made running the code of each paper over standard dataset (e.g. ImageNet, ...).

It is easy to check that PwC applies:

a) Taxonomizing of various technologies[21] by 5189 tasks[22]
b) Selection of characteristics[23] and measures
c) Definition of a quality rating module that includes 1 standard scenario[24]; benchmarks calculation; benchmark is intended a set of performance measurements (es. Top Accuracy-1) of all the available models against the same dataset in a limited simulation environment; measurements rating is displayed through a curve that links the performance score of each model over time (plane x=time, y=score, see fig. 4)
d) Evaluation (graphical) by each model and by each characteristic over a single scenario

It should be noted that:

- the simulation environment of PwC is simplified (same code language, single dataset,...) and cannot represent the production environment but gives immediate and homogeneous graphical comparisons among all the coded models solving a certain task, and
- the quality analysis is not complete as it is missing a multidimensional rating (e.g. a weighted score over multiple characteristics).

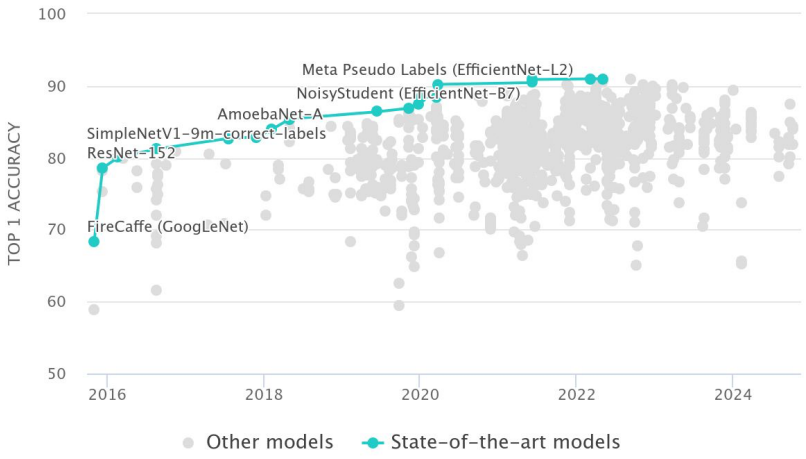

Figure 4: Historical benchmark for image classification

[19] In NLP applications, there is the general task of text classification, and among them there is the specific task for the machine to detect prompts with toxic text (e.g. biased questions, hate speech...)
[20] For the scope of this paper, we don't discuss the characteristic to which the measure of table 3 is referred; as hypothesis, it could be referred to Functional correctness.
[21] TechMethod of fig.1, it is a miscellaneous non homogeneous first-level description of the technological solution like NN, LLM,....
[22] Tasks correspond to algorithm in fig.1
[23] Most of scenarios refer to accuracy characteristic and its measures.
[24] Standard scenarios rely on standard dataset, both in training and input prompts. A set of scenarios is in general suitable for a set of measures (e.g. CivilComments and RealToxicityPrompts dataset are suitable for toxicity measurement).

Despite those limitations, steps a,b,c,d, are consistent with ISO approach, in conclusion both the measures and the evaluation method of PwC can be accounted as compliant to ISO AI quality standards.

7. Proposal

The proposal in this paper completes the proposal in [7][6]; there we showed how to account and represent measures from AI practices into the ISO/IEC 25000 format, here we explore how some existing quality evaluation practices are accountable as evaluation according AI ISO standards.

Finally, it is highlighted a difference between 'local' and 'global' benchmarking; the former it is always possible as it is always possible to measure and to rate on the quality target basis defined by the organization that handle the product as well; the latter, intended as a comparison among different products is challenging, due to the manifold of environments of the products even they are performing the same task.

8. Conclusion

The set of ISO AI standards clearly leverages SQuaRE product quality standards, both for measures and for evaluation. According SQuaRE and ISO/IEC 42001, it is always possible to define a quality model and a set of measures, even customized, to accomplish any measurements campaign.

At the same manner, it is always possible to define a quality rating module and accomplish an AI product evaluation on behalf of an organization or a third party.

Benchmarking is a useful method to rate, or to assign performance targets, or compare products, but is in general possible only locally, i.e., it is not possible for all the products, even if in the same category, for the limitations discussed above.

Such considerations are supported by the practices analyzed.

References

[1] D. Natale, A. Trenta, Examples of practical use of ISO/IEC 25000, Proceedings of IWESQ@APSEC 2019. URL: http://ceurws.org/Vol-2545/

[2] A. Trenta: ISO/IEC 25000 quality measures for A.I.: a geometrical approach, Proceedings of IWESQ@APSEC 2020. URL: http://ceur-ws.org/Vol-2800/

[3] A. Trenta, Data bias measurement: a geometrical approach through frames, Proceedings of IWESQ@APSEC 2021. URL: http://ceur-ws.org/Vol-3114/

[4] A. Simonetta, A. Trenta, M. C. Paoletti, and

[5] A. Vetrò, "Metrics for identifying bias in datasets," SYSTEM, 2021. [5] D. Natale, Extensions of ISO/IEC 25000 quality models to the context of Artificial Intelligence, Proceedings of IWESQ@APSEC 2022, URL: https://ceurws.org/Vol-3356

[6] A. Trenta: ISO/IEC 25000 and AI Product Quality Measurement Perspectives Proceedings IWESQ@APSEC 2022, URL: https://ceur-ws.org/Vol-3356

[7] A.Trenta "Accounting AI Measures as ISO/IEC 25000 Standards Measures", Proceedings of IWESQ@APSEC 2023, URL: http://ceur-ws.org/Vol-3612

[8] ISO/IEC 22989:2022 Information technology — Artificial intelligence Artificial intelligence concepts and terminology. URL: https://www.iso.org/standard/74296.html

[9] ISO/IEC 23053:2022 - Framework for Artificial Intelligence (AI) - Systems Using Machine Learning (ML) URL: https://www.iso.org/standard/74438.html

[10] ISO/IEC TR 24372 Information technology — Artificial intelligence — Overview of computational approaches for AI systems URL: https://www.iso.org/standard/78508.html

[11] ISO/IEC TR 24030:2024 Information technology — Artificial intelligence (AI) Use cases URL: https://www.iso.org/standard/84144.html

[12] ISO/IEC DIS 25059 Software engineering Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality Model for AI-based systems. URL: https://www.iso.org/standard/80655.html

[13] ISO/IEC 42001 Information technology Artificial intelligence — Management system. URL: https://www.iso.org/standard/81230.html

[14] ISO/IEC TS 25058 Software engineering Systems and software Quality Requirements and Evaluation (SquaRE) — Guidance for quality evaluation of artificial intelligence (AI) systems. URL: https://www.iso.org/standard/82570.html

[15] ISO/IEC CD TS 17847 -Information Technology — Artificial Intelligence Verification and validation analysis of AI systems

[16] ISO/IEC 25040 - Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) Quality evaluation framework

[17] ISO/IEC AWI TR 42106 Information technology — Artificial intelligence Overview of differentiated benchmarking of AI system quality characteristics

[18] ISO/IEC DIS 25002 Systems and Software engineering - Systems and software Quality

Requirements and Evaluation (SQuaRE) Quality models overview and usage
[19] ISO/IEC 5259-2 Artificial intelligence Data quality for analytics and ML — Part 2: Data quality measures. URL: https://www.iso.org/standard/81860.html
[20] ISO/IEC TR 24029-1:2021 Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview
[21] ISO/IEC 24029-2:2023 Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods
[22] ISO/IEC TR 12182:2015 - Systems and software engineering — Framework for categorization of IT systems and software, and guide for applying it, URL: https://www.iso.org/standard/63611.html
[23] P.Liang, R. Bommasani, T. Lee et al., Holistic Evaluation of Language Models, Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford University, 2022
[24] ITU-T F.748.11 Metrics and evaluation methods for a deep neural network processor benchmark, 2020
[25] ITU-T F.748.12 Deep learning software framework evaluation methodology, 2021
[26] Meta Research, Papers with Code resource. URL: https://paperswithcode.com/sota