
Umetna inteligenca (UI) - Metode vrednotenja za natančne sisteme za obdelavo naravnega jezika (ISO/IEC DIS 23282:2026)

Artificial Intelligence - Evaluation methods for accurate natural language processing systems (ISO/IEC DIS 23282:2026)

Evaluationsmethoden für die Genauigkeit von Systemen zur Verarbeitung natürlicher Sprache (ISO/IEC DIS 23282:2026)

Intelligence artificielle - Méthodes d'évaluation pour des systèmes de traitement automatique du langage naturel performants (ISO/IEC DIS 23282:2026)

Ta slovenski standard je istoveten z: prEN ISO/IEC 23282

ICS:

35.240.01	Uporabniške rešitve informacijske tehnike in tehnologije na splošno	Application of information technology in general
-----------	---	--

oSIST prEN ISO/IEC 23282:2026

en,fr,de

Sample Document

get full document from standards.iteh.ai



DRAFT International Standard

ISO/IEC DIS 23282

Artificial Intelligence — Evaluation methods for accurate natural language processing systems

*Intelligence artificielle — Méthodes d'évaluation pour des
systèmes de traitement automatique du langage naturel
performants*

ICS: 35.240.01

ISO/IEC JTC 1/SC 42

Secretariat: **ANSI**

Voting begins on:
2026-09-02

Voting terminates on:
2026-11-25

Member bodies are requested to consult relevant national interests in ISO/TC 37 before casting their ballot to the e-Balloting application.

This document has not been edited by the ISO Central Secretariat.

ISO/CEN PARALLEL PROCESSING

Reference number
ISO/IEC DIS 23282:2026(en)

THIS DOCUMENT IS A DRAFT CIRCULATED FOR COMMENTS AND APPROVAL. IT IS THEREFORE SUBJECT TO CHANGE AND MAY NOT BE REFERRED TO AS AN INTERNATIONAL STANDARD UNTIL PUBLISHED AS SUCH.

IN ADDITION TO THEIR EVALUATION AS BEING ACCEPTABLE FOR INDUSTRIAL, TECHNOLOGICAL, COMMERCIAL AND USER PURPOSES, DRAFT INTERNATIONAL STANDARDS MAY ON OCCASION HAVE TO BE CONSIDERED IN THE LIGHT OF THEIR POTENTIAL TO BECOME STANDARDS TO WHICH REFERENCE MAY BE MADE IN NATIONAL REGULATIONS.

RECIPIENTS OF THIS DRAFT ARE INVITED TO SUBMIT, WITH THEIR COMMENTS, NOTIFICATION OF ANY RELEVANT PATENT RIGHTS OF WHICH THEY ARE AWARE AND TO PROVIDE SUPPORTING DOCUMENTATION.

© ISO/IEC 2026

ISO/IEC DIS 23282:2026(en)

Sample Document

get full document from standards.iteh.ai



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2026

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

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

© ISO/IEC 2026 – All rights reserved

ISO/IEC DIS 23282:2026(en)

Contents

Page

Foreword	vii
Introduction	viii
1 Scope	1
2 Normative references	1
3 Terms and definitions	1
4 Abbreviations	4
5 Overview	5
5.1 General	5
5.2 Types of evaluation methods	5
5.3 Contents of this document	6
6 Description of evaluation methods	7
6.1 General	7
6.2 Automated metrics	7
6.2.1 General	7
6.2.2 Accuracy, precision and recall, F1 score	8
6.2.3 Precision @ K, recall @ K, F1 @ K	8
6.2.4 Purity, inverse purity	8
6.2.5 Edit distance metrics	8
6.2.6 Entropy, mutual information and normalized mutual information	9
6.2.7 N-gram and subsequence matching metrics	10
6.2.8 Embeddings-based metrics	13
6.2.9 Learned metrics	14
6.2.10 EER (equal error rate)	16
6.2.11 Jaccard coefficient, Rand score, adjusted Rand score, Fowlkes-Mallows index	16
6.2.12 Silhouette coefficient, silhouette index, Davies-Bouldin index, Calinski-Harabasz index	16
6.2.13 B ³	17
6.2.14 AER (alignment error rate)	18
6.2.15 METEOR	18
6.2.16 Perplexity	18
6.2.17 DCF, minDCF and actDCF	18
6.3 Human evaluation protocols	19
6.3.1 General	19
6.3.2 Types of human evaluation protocols	20
6.3.3 MOS (mean opinion score)	21
6.3.4 DRT (diagnostic rhyme test)	21
6.3.5 MRT (modified rhyme test)	21
6.3.6 MUSHRA	22
6.4 Multi-dimensional analysis criteria	22
6.4.1 General	22
6.4.2 Quality criteria	22
6.4.3 Error criteria	23
7 Shared evaluation methods per task prototype	24
7.1 General	24
7.2 Categorisation	24
7.2.1 General	24
7.2.2 Automated metric: accuracy	24
7.2.3 Automated metrics: precision, recall, F1-score	24
7.3 Identification	25
7.3.1 General	25
7.3.2 Automated metric: accuracy	25
7.3.3 Automated metrics: precision, recall, F1-score	26

ISO/IEC DIS 23282:2026(en)

7.3.4	Automated metric: EER.....	26
7.3.5	Automated metrics: DCF, minDCF, actDCF	26
7.4	Segmentation	26
7.4.1	General	26
7.4.2	Automated metrics: precision, recall, F1-score	27
7.5	Sequence labelling.....	28
7.5.1	General	28
7.5.2	Automated metric: accuracy.....	28
7.5.3	Automated metrics: precision, recall, F1-score	28
7.6	Per-element identification.....	29
7.6.1	General	29
7.6.2	Automated metric: accuracy.....	29
7.6.3	Automated metrics: precision, recall, F1-score	29
7.7	Set creation.....	30
7.7.1	General	30
7.7.2	Automated metrics: silhouette coefficient, silhouette index, Davies-Bouldin index, Calinski-Harabasz index.....	30
7.7.3	Automated metrics: purity, inverse purity.....	30
7.7.4	Automated metrics: Jaccard coefficient, Rand score, adjusted Rand score, Fowlkes-Mallows index.....	30
7.7.5	Automated metrics: mutual information, normalized mutual information	31
7.7.6	Automated metric: MUC.....	31
7.7.7	Automated metric: B ³	32
7.8	Ranking.....	32
7.8.1	General	32
7.8.2	Automated metrics: precision @ K, recall @ K, F1 @ K.....	32
7.9	Sequence generation.....	32
7.9.1	General	32
7.9.2	Automated metric: BLEU	33
7.9.3	Automated metrics: ROUGE-1, ROUGE-2, ROUGE-n.....	33
7.9.4	Automated metric: ROUGE-L.....	33
7.9.5	Automated metric: perplexity.....	33
7.9.6	Automated metrics: CER, WER, PER.....	34
7.9.7	Automated metrics: BERTscore	34
7.9.8	Automated metrics: COMET regression metric, COMET ranking metric	34
7.10	Structure generation.....	34
8	Evaluation methods per task.....	35
8.1	General.....	35
8.2	Tasks related to shallow analysis of natural language contents	35
8.2.1	General.....	35
8.2.2	Language identification	35
8.2.3	Sentence boundary detection.....	37
8.2.4	Tokenization.....	37
8.2.5	Lemmatization	38
8.2.6	Part-of-speech tagging.....	39
8.2.7	Syntactic parsing.....	40
8.2.8	Morphological analysis.....	40
8.2.9	Word sense disambiguation.....	40
8.2.10	Coreference resolution.....	43
8.2.11	Anaphora resolution	46
8.2.12	Discourse parsing.....	47
8.2.13	Document segmentation	47
8.2.14	Parallel corpora alignment	48
8.2.15	Speech activity detection	48
8.2.16	Utterance boundary detection.....	49
8.2.17	Phrase boundary detection	49
8.2.18	Speaker change detection	50
8.2.19	Speaker diarization.....	50

ISO/IEC DIS 23282:2026(en)

8.3	Tasks related to author identification and profiling.....	50
8.3.1	General.....	50
8.3.2	Authorship attribution.....	50
8.3.3	Speaker recognition.....	52
8.3.4	Handwriting identification.....	55
8.3.5	Style analysis.....	57
8.3.6	Speaker role identification.....	57
8.3.7	Speaker gender identification.....	58
8.3.8	Impersonation detection.....	59
8.4	Tasks related to document analysis and mining.....	59
8.4.1	General.....	59
8.4.2	Topic analysis.....	59
8.4.3	Named entity recognition.....	60
8.4.4	Entity linking.....	61
8.4.5	Relation extraction.....	62
8.4.6	Slot filling.....	62
8.4.7	Claim extraction and fact extraction.....	62
8.4.8	Sentiment analysis.....	63
8.4.9	Emotion recognition.....	64
8.4.10	User intent detection.....	65
8.4.11	Author intention recognition.....	66
8.4.12	Offensive speech detection.....	67
8.5	Tasks related to corpus-level analysis and mining.....	68
8.5.1	General.....	68
8.5.2	Document recommendation.....	68
8.5.3	Content-based document search.....	68
8.5.4	Semantic clustering.....	69
8.5.5	Thematic clustering.....	69
8.6	Tasks related to semantics and meaning.....	70
8.6.1	General.....	70
8.6.2	Automatic term extraction.....	70
8.6.3	Word sense induction.....	70
8.6.4	Term set expansion.....	71
8.6.5	Semantic relation extraction.....	71
8.6.6	Embeddings.....	71
8.6.7	Representation learning.....	71
8.6.8	Language modelling.....	72
8.7	Tasks related with questions and dialogue.....	72
8.7.1	General.....	72
8.7.2	Question answering.....	72
8.7.3	Generation of multimedia contents from prompts.....	74
8.8	Tasks involving the generation of linguistic contents.....	74
8.8.1	General.....	74
8.8.2	Automatic summarization.....	74
8.8.3	Machine translation.....	75
8.8.4	Document denoising.....	77
8.8.5	Language error correction.....	78
8.8.6	Text normalization.....	79
8.8.7	Sensitive data treatment.....	79
8.8.8	Voice anonymization.....	79
8.8.9	Description generation.....	79
8.8.10	Vanilla text generation.....	80
8.8.11	Constrained text generation.....	82
8.8.12	Captioning.....	84
8.8.13	Source code generation.....	84
8.8.14	Source code documentation.....	84
8.8.15	Paraphrase generation.....	85
8.8.16	Question generation.....	85
8.8.17	Distractor generation.....	85

ISO/IEC DIS 23282:2026(en)

8.9	Tasks involving conversions among modalities.....	85
8.9.1	General.....	85
8.9.2	Automatic speech recognition	86
8.9.3	Text-to-speech synthesis	86
8.9.4	Optical character recognition.....	88
8.9.5	Lip reading.....	89
8.10	Transversal functionalities	89
8.10.1	General.....	89
8.10.2	Domain adaptation.....	89
9	Evaluation methodology for multi-component NLP systems	90
10	Requirements on resources.....	91
10.1	General.....	91
10.2	Requirements on test data.....	91
10.3	Requirements on computing resources.....	91
	Bibliography.....	92

Sample Document

get full document from standards.iteh.ai

ISO/IEC DIS 23282:2026(en)**Foreword**

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

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 ISO documents 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).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

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

For an explanation on 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 the following URL: www.iso.org/iso/foreword.html.

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.

get full document from standards.iteh.ai

ISO/IEC DIS 23282:2026(en)**Introduction**

This document focuses on the evaluation of natural language processing (NLP) systems, as regards their functional suitability as defined in ISO/IEC 25059:2023 [1]. This includes functional correctness (often informally referred to as 'performance' by NLP practitioners, which ISO/IEC 25059:2023 [1], Annex C discourages due to ambiguity, or 'accuracy' in some regulation, which can be inconsistent with technical usage), as well as other quality characteristics such as functional adaptability.

While other standards exist on the evaluation of AI systems, such as ISO/IEC DIS 4213, they only contain generic metrics, targeting the evaluation of some base AI tasks like classification, and those metrics are either not sufficiently specified to be applied in turn on NLP systems or not applicable to part of these NLP systems.

This document therefore builds on ISO/IEC DIS 4213 to provide the NLP-specific complements that are necessary to enable the evaluation of NLP systems in particular. This includes additional specifications on generic metrics used beyond NLP, as well as NLP-specific evaluation methods, encompassing both automated and human evaluation.

Sample Document

get full document from standards.iteh.ai

Artificial Intelligence — Evaluation methods for accurate natural language processing systems

1 Scope

This document specifies the evaluation of natural language processing systems, in the sense of measuring the quality of a system's results to assess its functional suitability. It provides a definition of evaluation methods for those systems, together with guidance on how to select, implement and interpret those evaluation methods. This document covers quantitative metrics as well as other evaluation methods. It includes requirements on the implementation of the described metrics, and further requirements on the technical resources involved in the evaluation process.

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 DIS 4213, *Information technology — Artificial Intelligence — Performance measurement for AI classification, regression, clustering and recommendation tasks*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

- IEC Electropedia: available at <http://www.electropedia.org/>
- ISO Online browsing platform: available at <http://www.iso.org/obp>

3.1

artificial intelligence, noun

AI, noun

(discipline) research and development of mechanisms and applications of AI systems

Note 1 to entry: Research and development can take place across any number of fields such as computer science, data science, humanities, mathematics and natural sciences.

[SOURCE: ISO/IEC 22989:2022 [2], 3.1.3]

3.2

natural language, noun

language that is or was in active use in a community of people and whose rules are deduced from usage

Note 1 to entry: Natural language is any human language, which can be expressed in text, speech, sign language, etc.

Note 2 to entry: Natural language is any human language, such as English, Spanish, Arabic, Chinese or Japanese, to be distinguished from programming and formal languages, such as Java, Fortran, C++ or First-Order Logic.

[SOURCE: ISO/IEC 22989:2022 [2], 3.6.7]

ISO/IEC DIS 23282:2026(en)

3.3

natural language processing, noun**NLP**, noun(discipline) discipline concerned with the way systems acquire, process and interpret *natural language* (3.2)

[SOURCE: ISO/IEC 22989:2022 [2], 3.6.10]

3.4

natural language processing, noun**NLP**, noun(system) information processing based upon *natural language understanding* (3.5) or *natural language generation* (3.6)

[SOURCE: ISO/IEC 22989:2022 [2], 3.6.9]

3.5

natural language understanding, noun**NLU**, noun

natural language comprehension, noun

extraction of information, by a functional unit, from text or speech communicated to it in a *natural language* (3.2), and the production of a description for both the given text or speech and what it represents

[SOURCE: ISO/IEC 22989:2022 [2], 3.6.11]

3.6

natural language generation, noun**NLG**, nountask of converting data carrying semantics into *natural language* (3.2)

[SOURCE: ISO/IEC 22989:2022 [2], 3.6.8]

3.7

multilingual, noun(system) able to process inputs in more than one *natural language* (3.2)

Note 1 to entry: Individual inputs can be in a single language or contain multiple languages.

3.8

knowledge, noun

(artificial intelligence) abstracted information about objects, events, concepts or rules, their relationships and properties, organized for goal-oriented systematic use

Note 1 to entry: Knowledge in the *AI* (3.1) domain does not imply a cognitive capability, contrary to usage of the term in some other domains. In particular, knowledge does not imply the cognitive act of understanding.

Note 2 to entry: Information can exist in numeric or symbolic form.

Note 3 to entry: Information is data that has been contextualized, so that it is interpretable. Data is created through abstraction or measurement from the world.

[SOURCE: ISO/IEC 22989:2022 [2], 3.1.21]

3.9

prediction, noun

primary output of an AI system when provided with input data or information

Note 1 to entry: Predictions can be followed by additional outputs, such as recommendations, decisions and actions.

Note 2 to entry: Prediction does not necessarily refer to predicting something in the future.

Note 3 to entry: Predictions can refer to various kinds of data analysis or production applied to new data or historical data (including translating text, creating synthetic images or diagnosing a previous power failure).

ISO/IEC DIS 23282:2026(en)

[SOURCE: ISO/IEC 22989:2022 [2], 3.1.27]

3.10

functional correctness, noun

degree to which a product or system provides the correct results with the needed degree of precision

Note 1 to entry: AI systems, and particularly those using *machine learning models* (3.14), do not usually provide functional correctness in all observed circumstances.

Note 2 to entry: Functional correctness is part of *functional suitability* (3.11).

[SOURCE: ISO/IEC 25059:2023 [1], 3.2.3, modified — added Note 2]

3.11

functional suitability, noun

capability of a product to provide functions that meet stated and implied needs of intended users when it is used under specified conditions

Note 1 to entry: Functional suitability is concerned with whether the functions meet not only stated and implied needs, but also the functional specification.

[SOURCE: ISO/IEC 25010:2023 [3], 3.1]

3.12

test data, noun

evaluation data, noun

data used to assess the performance of a final model

[SOURCE: ISO/IEC 22989:2022 [2], 3.2.14, modified — removed Note 1]

3.13

independent, adj

statistically independent, adj

(dataset) property of a dataset where no characteristic is shared with another dataset other than having the same statistical distribution

Note 1 to entry: This concept of independence originates from probability theory, where it refers to the property of two events whereby the occurrence of one event does not affect the probability of occurrence of the other.

Note 2 to entry: Independent datasets sharing the same statistical distribution are said to be independent and identically distributed.

Note 3 to entry: Independent datasets are not necessarily disjoint, provided that the shared elements are the result of having the same distribution.

Note 4 to entry: Datasets that are not independent can lead to the appearance of generalisation where none exists or otherwise invalidate evaluation results derived from those datasets.

Note 5 to entry: Unintended violation of independence resulting in inflated or invalid evaluation results is referred to as data leakage.

3.14

machine learning model, noun

mathematical construct that generates an inference or *prediction* (3.9) based on input data or information

EXAMPLE If a univariate linear function ($y = \theta_0 + \theta_1 x$) has been trained using linear regression, the resulting model can be $y = 3 + 7x$.

Note 1 to entry: A machine learning model results from training based on a machine learning algorithm.

[SOURCE: ISO/IEC 22989:2022 [2], 3.3.7]

ISO/IEC DIS 23282:2026(en)

3.15**bias**, noun

systematic difference in treatment of certain objects, people or groups in comparison to others

Note 1 to entry: Treatment is any kind of action, including perception, observation, representation, *prediction* ([3.9](#)) or decision.

[SOURCE: ISO/IEC 22989:2022 [\[2\]](#), 3.5.4]

3.16**automatic**, adj**automation**, noun**automated**, adj

pertaining to a process or system that, under specified conditions, functions without human intervention

[SOURCE: ISO/IEC 22989:2022 [\[2\]](#), 3.1.7]

4 Abbreviations

AER	alignment error rate
B ³	B-cubed
BERT	bidirectional encoder representations from transformers
BLEU	bilingual evaluation understudy
CER	character error rate
COMET	crosslingual optimized metric for evaluation of translation
DCF	detection cost function
DRT	diagnostic rhyme test
EER	equal error rate
METEOR	metric for evaluation of translation with explicit ordering
ML	machine learning
MOS	mean opinion score
MRT	modified rhyme test
MUC	message understanding conference
MUSHRA	Multiple Stimuli with Hidden Reference and Anchor
PER	phoneme error rate
ROUGE	recall-oriented understudy for gisting evaluation
WER	word error rate

ISO/IEC DIS 23282:2026(en)

5 Overview

5.1 General

Compared to evaluation practices for other AI systems, natural language processing (NLP) has its own specificities. While the main principles and caveats remain, the concrete practices, and in particular the evaluation methods themselves, can be substantially affected in NLP, both due to the inherent constraints and variability associated with processing natural language content (text, speech...) and to existing practices by NLP practitioners.

This document approaches the specificities of NLP evaluation from the angle of tasks. The task is a suitable way to formalize the processing performed by the AI system, at a level of abstraction that enables to capture the methodological specificities of certain types of processing, while not delving into the specific considerations of the use case and the exact data that is being processed.

EXAMPLE AI systems performing a classification task can be evaluated through metrics like precision, recall or F1, as defined and specified in ISO/IEC DIS 4213, 6. However, this is not applicable to machine translation or automatic speech recognition and different evaluation methods are used for such systems. Cases like named entity recognition are in between: evaluation metrics like precision, recall or F1 are applicable, but in a different way from how they are applied for classification, and which warrants further specification. For instance, that specification can include determining whether the metric is to be applied on token-level labels or on the entity spans, or it can establish the specific criteria for counting a given predicted entity span as being a true positive.

This document should therefore be read and used in conjunction with ISO/IEC TR 23281 [4]. The description of the various tasks addressed in this document can be found in ISO/IEC TR 23281 [4], 7, and this document focuses on which evaluation methods can be applied to each task and how.

It is complementary to ISO/IEC DIS 4213, which addresses evaluation methods for classification, regression, clustering and recommendation, and to similar standards focusing on evaluation in other fields of AI, such as computer vision. Evaluation methods from ISO/IEC DIS 4213 are built upon by some evaluation methods in this document, and evaluation methods from computer vision are typically combined with the present ones, either in the context of multi-modal processing (for which evaluating the processing on both aspects is warranted) or due to the joint presence of NLP components and Computer Vision components within the same AI system.

5.2 Types of evaluation methods

The evaluation methods targeted by this document encompass both automated and human evaluation:

- Automated metrics produce a quantification of the quality of the NLP system's outputs, by performing automatic computations on a test dataset (usually annotated with reference expected outputs) and on the outputs produced by the NLP system when processing the inputs of that test dataset. These computations can include statistical measurements but also more complex processes such as involving another NLP system to perform the computation.
- Human evaluation protocols aim at measuring the human judgement on the NLP system's outputs, in a way that is as reproducible as possible. It relies on a precise definition of the conditions in which this human judgement is collected from a pool of human subjects, as well as a set of quality controls that can be applied on the evaluation process, which can in turn involve statistical measurements for aggregation or filtering purposes.

Both types of evaluation methods are typically complementary.

Each evaluation method has its own limitations and typically focuses on only certain aspects of the output's quality (e.g. accounting only for certain types of errors, or penalizing them differently). But, in general, a given evaluation method produces a single value, meant to quantify that quality. A typical strategy to mitigate these limitations is to apply multiple evaluation methods in a complementary manner, selected based on the use case needs.