

[ ![Logo](../../\_static/logo.png) ](../../index.html)

## Getting Started

- \* [Installation](../../installation.html)
- \* [Install with pip](../../installation.html#install-with-pip)
- \* [Install with Conda](../../installation.html#install-with-conda)
- \* [Install from Source](../../installation.html#install-from-source)
- \* [Editable Install](../../installation.html#editable-install)
- \* [Install PyTorch with CUDA support](../../installation.html#install-pytorch-with-cuda-support)
- \* [Quickstart](../../quickstart.html)
- \* [Sentence Transformer](../../quickstart.html#sentence-transformer)
- \* [Cross Encoder](../../quickstart.html#cross-encoder)
- \* [Next Steps](../../quickstart.html#next-steps)

## Sentence Transformer

- \* [Usage](../../sentence\_transformer/usage/usage.html)
- \* [Computing Embeddings](../../examples/applications/computing-embeddings/README.html)
  - \* [Initializing a Sentence Transformer Model](../../examples/applications/computing-embeddings/README.html#initializing-a-sentence-transformer-model)
  - \* [Calculating Embeddings](../../examples/applications/computing-embeddings/README.html#calculating-embeddings)
  - \* [Prompt Templates](../../examples/applications/computing-embeddings/README.html#prompt-templates)

[Length\]\(../../examples/applications/computing-embeddings/README.html#id1\)](#)

[\\* \[Multi-Process / Multi-GPU Encoding\]\(../../examples/applications/computing-embeddings/README.html#multi-process-multi-gpu-encoding\)](#)

[\\* \[Semantic Textual Similarity\]\(../../sentence\\_transformer/usage/semantic\\_textual\\_similarity.html\)](#)

[\\* \[Similarity Calculation\]\(../../sentence\\_transformer/usage/semantic\\_textual\\_similarity.html#similarity-calculation\)](#)

[\\* \[Semantic Search\]\(../../examples/applications/semantic-search/README.html\)](#)

[\\* \[Background\]\(../../examples/applications/semantic-search/README.html#background\)](#)

[\\* \[Symmetric vs. Asymmetric Semantic Search\]\(../../examples/applications/semantic-search/README.html#symmetric-vs-asymmetric-semantic-search\)](#)

[\\* \[Manual Implementation\]\(../../examples/applications/semantic-search/README.html#manual-implementation\)](#)

[\\* \[Optimized Implementation\]\(../../examples/applications/semantic-search/README.html#optimized-implementation\)](#)

[\\* \[Speed Optimization\]\(../../examples/applications/semantic-search/README.html#speed-optimization\)](#)

[\\* \[Elasticsearch\]\(../../examples/applications/semantic-search/README.html#elasticsearch\)](#)

[\\* \[Approximate Nearest Neighbor\]\(../../examples/applications/semantic-search/README.html#approximate-nearest-neighbor\)](#)

[\\* \[Retrieve & Re-Rank\]\(../../examples/applications/semantic-search/README.html#retrieve-re-rank\)](#)

\* [Examples](../../examples/applications/semantic-search/README.html#examples)

\* [Retrieve & Re-Rank](../../examples/applications/retrieve\_rerank/README.html)

\* [Retrieve & Re-Rank

Pipeline](../../examples/applications/retrieve\_rerank/README.html#retrieve-re-rank-pipeline)

\* [Retrieval:

Bi-Encoder](../../examples/applications/retrieve\_rerank/README.html#retrieval-bi-encoder)

\* [Re-Ranker:

Cross-Encoder](../../examples/applications/retrieve\_rerank/README.html#re-ranker-cross-encoder)

\* [Example Scripts](../../examples/applications/retrieve\_rerank/README.html#example-scripts)

\* [Pre-trained Bi-Encoders

(Retrieval)](../../examples/applications/retrieve\_rerank/README.html#pre-trained-bi-encoders-retrieval)

\* [Pre-trained Cross-Encoders

(Re-Ranker)](../../examples/applications/retrieve\_rerank/README.html#pre-trained-cross-encoders-re-ranker)

\* [Clustering](../../examples/applications/clustering/README.html)

\* [k-Means](../../examples/applications/clustering/README.html#k-means)

\* [Agglomerative

Clustering](../../examples/applications/clustering/README.html#agglomerative-clustering)

\* [Fast Clustering](../../examples/applications/clustering/README.html#fast-clustering)

\* [Topic Modeling](../../examples/applications/clustering/README.html#topic-modeling)

\* [Paraphrase Mining](../../examples/applications/paraphrase-mining/README.html)

\*

[`paraphrase\_mining()`](../../examples/applications/paraphrase-mining/README.html#sentence\_transformers.util.paraphrase\_mining)

\* [Translated Sentence

Mining](../../examples/applications/parallel-sentence-mining/README.html)

\* [Margin Based

Mining](../../examples/applications/parallel-sentence-mining/README.html#margin-based-mining)

\* [Examples](../../examples/applications/parallel-sentence-mining/README.html#examples)

\* [Image Search](../../examples/applications/image-search/README.html)

\* [Installation](../../examples/applications/image-search/README.html#installation)

\* [Usage](../../examples/applications/image-search/README.html#usage)

\* [Examples](../../examples/applications/image-search/README.html#examples)

\* [Embedding Quantization](../../examples/applications/embedding-quantization/README.html)

\* [Binary

Quantization](../../examples/applications/embedding-quantization/README.html#binary-quantization)

\* [Scalar (int8)

Quantization](../../examples/applications/embedding-quantization/README.html#scalar-int8-quantization)

\* [Additional

extensions](../../examples/applications/embedding-quantization/README.html#additional-extensions)

\* [Demo](../../examples/applications/embedding-quantization/README.html#demo)

\* [Try it

yourself](../../examples/applications/embedding-quantization/README.html#try-it-yourself)

\* [Speeding up Inference](../../sentence\_transformer/usage/efficiency.html)

\* [PyTorch](../../sentence\_transformer/usage/efficiency.html#pytorch)

\* [ONNX](../../sentence\_transformer/usage/efficiency.html#onnx)

\* [OpenVINO](../../sentence\_transformer/usage/efficiency.html#openvino)

\* [Benchmarks](../../sentence\_transformer/usage/efficiency.html#benchmarks)

\* [Creating Custom Models](../../sentence\_transformer/usage/custom\_models.html)

\* [Structure of Sentence Transformer

Models](../../sentence\_transformer/usage/custom\_models.html#structure-of-sentence-transformer-models)

\* [Sentence Transformer Model from a Transformers

Model](../../sentence\_transformer/usage/custom\_models.html#sentence-transformer-model-from-a-transformers-model)

\* [Pretrained Models](../../sentence\_transformer/pretrained\_models.html)

\* [Original Models](../../sentence\_transformer/pretrained\_models.html#original-models)

\* [Semantic Search

Models](../../sentence\_transformer/pretrained\_models.html#semantic-search-models)

\* [Multi-QA Models](../../sentence\_transformer/pretrained\_models.html#multi-qa-models)

\* [MSMARCO Passage

Models](../../sentence\_transformer/pretrained\_models.html#msmarco-passage-models)

\* [Multilingual Models](../../sentence\_transformer/pretrained\_models.html#multilingual-models)

\* [Semantic Similarity

Models](../../sentence\_transformer/pretrained\_models.html#semantic-similarity-models)

\* [Bitext Mining](../../sentence\_transformer/pretrained\_models.html#bitext-mining)

\* [Image & Text-Models](../../sentence\_transformer/pretrained\_models.html#image-text-models)

\* [INSTRUCTOR models](../../sentence\_transformer/pretrained\_models.html#instructor-models)

\* [Scientific Similarity

Models](../../sentence\_transformer/pretrained\_models.html#scientific-similarity-models)

\* [Training Overview](../../sentence\_transformer/training\_overview.html)

\* [Why Finetune?](../../sentence\_transformer/training\_overview.html#why-finetune)

\* [Training Components](../../sentence\_transformer/training\_overview.html#training-components)

\* [Dataset](../../sentence\_transformer/training\_overview.html#dataset)

\* [Dataset Format](../../sentence\_transformer/training\_overview.html#dataset-format)

\* [Loss Function](../../sentence\_transformer/training\_overview.html#loss-function)

- \* [Training Arguments](../../sentence\_transformer/training\_overview.html#training-arguments)
- \* [Evaluator](../../sentence\_transformer/training\_overview.html#evaluator)
- \* [Trainer](../../sentence\_transformer/training\_overview.html#trainer)
- \* [Callbacks](../../sentence\_transformer/training\_overview.html#callbacks)
- \* [Multi-Dataset Training](../../sentence\_transformer/training\_overview.html#multi-dataset-training)
- \* [Deprecated Training](../../sentence\_transformer/training\_overview.html#deprecated-training)
- \* [Best Base Embedding Models](../../sentence\_transformer/training\_overview.html#best-base-embedding-models)
- \* [Dataset Overview](../../sentence\_transformer/dataset\_overview.html)
  - \* [Datasets on the Hugging Face Hub](../../sentence\_transformer/dataset\_overview.html#datasets-on-the-hugging-face-hub)
  - \* [Pre-existing Datasets](../../sentence\_transformer/dataset\_overview.html#pre-existing-datasets)
- \* [Loss Overview](../../sentence\_transformer/loss\_overview.html)
  - \* [Loss modifiers](../../sentence\_transformer/loss\_overview.html#loss-modifiers)
  - \* [Distillation]
    - \* [Commonly used Loss Functions](../../sentence\_transformer/loss\_overview.html#commonly-used-loss-functions)
    - \* [Custom Loss Functions](../../sentence\_transformer/loss\_overview.html#custom-loss-functions)
- \* [Training Examples](../../sentence\_transformer/training/examples.html)
  - \* [Semantic Textual Similarity](../../examples/training/sts/README.html)
    - \* [Training data](../../examples/training/sts/README.html#training-data)
    - \* [Loss Function](../../examples/training/sts/README.html#loss-function)
  - \* [Natural Language Inference](../../examples/training/nli/README.html)
    - \* [Data](../../examples/training/nli/README.html#data)
    - \* [SoftmaxLoss](../../examples/training/nli/README.html#softmaxloss)
    - \* [MultipleNegativesRankingLoss](../../examples/training/nli/README.html#multiplenegativesrankin

gloss)

- \* [Paraphrase Data](../../examples/training/paraphrases/README.html)
- \* [Pre-Trained Models](../../examples/training/paraphrases/README.html#pre-trained-models)
- \* [Quora Duplicate Questions](../../examples/training/quora\_duplicate\_questions/README.html)
- \* [Training](../../examples/training/quora\_duplicate\_questions/README.html#training)

\*

[MultipleNegativesRankingLoss](../../examples/training/quora\_duplicate\_questions/README.html#multiplenegativesrankingloss)

\*

[Pretrained

Models](../../examples/training/quora\_duplicate\_questions/README.html#pretrained-models)

- \* [MS MARCO](../../examples/training/ms\_marco/README.html)
- \* [Bi-Encoder](../../examples/training/ms\_marco/README.html#bi-encoder)
- \* [Matryoshka Embeddings](../../examples/training/matryoshka/README.html)
- \* [Use Cases](../../examples/training/matryoshka/README.html#use-cases)
- \* [Results](../../examples/training/matryoshka/README.html#results)
- \* [Training](../../examples/training/matryoshka/README.html#training)
- \* [Inference](../../examples/training/matryoshka/README.html#inference)
- \* [Code Examples](../../examples/training/matryoshka/README.html#code-examples)
- \* [Adaptive Layers](../../examples/training/adaptive\_layer/README.html)
- \* [Use Cases](../../examples/training/adaptive\_layer/README.html#use-cases)
- \* [Results](../../examples/training/adaptive\_layer/README.html#results)
- \* [Training](../../examples/training/adaptive\_layer/README.html#training)
- \* [Inference](../../examples/training/adaptive\_layer/README.html#inference)
- \* [Code Examples](../../examples/training/adaptive\_layer/README.html#code-examples)
- \* [Multilingual Models](../../examples/training/multilingual/README.html)

\*

[Extend your own

models](../../examples/training/multilingual/README.html#extend-your-own-models)

- \* [Training](../../examples/training/multilingual/README.html#training)
- \* [Datasets](../../examples/training/multilingual/README.html#datasets)
  - \* [Sources for Training Data](../../examples/training/multilingual/README.html#sources-for-training-data)
  - \* [Evaluation](../../examples/training/multilingual/README.html#evaluation)
    - \* [Available Pre-trained Models](../../examples/training/multilingual/README.html#available-pre-trained-models)
    - \* [Usage](../../examples/training/multilingual/README.html#usage)
    - \* [Performance](../../examples/training/multilingual/README.html#performance)
    - \* [Citation](../../examples/training/multilingual/README.html#citation)
  - \* [Model Distillation](../../examples/training/distillation/README.html)
    - \* [Knowledge Distillation](../../examples/training/distillation/README.html#knowledge-distillation)
      - \* [Speed - Performance Trade-Off](../../examples/training/distillation/README.html#speed-performance-trade-off)
      - \* [Dimensionality Reduction](../../examples/training/distillation/README.html#dimensionality-reduction)
    - \* [Quantization](../../examples/training/distillation/README.html#quantization)
  - \* [Augmented SBERT](../../examples/training/data\_augmentation/README.html)
    - \* [Motivation](../../examples/training/data\_augmentation/README.html#motivation)
      - \* [Extend to your own datasets](../../examples/training/data\_augmentation/README.html#extend-to-your-own-datasets)
      - \* [Methodology](../../examples/training/data\_augmentation/README.html#methodology)
        - \* [Scenario 1: Limited or small annotated datasets (few labeled sentence-pairs)](../../examples/training/data\_augmentation/README.html#scenario-1-limited-or-small-annotated-datasets-few-labeled-sentence-pairs)
        - \* [Scenario 2: No annotated datasets (Only unlabeled



sentence-pairs)](../../examples/training/data\_augmentation/README.html#scenario-2-no-annotated-datasets-only-unlabeled-sentence-pairs)

- \* [Training](../../examples/training/data\_augmentation/README.html#training)
- \* [Citation](../../examples/training/data\_augmentation/README.html#citation)
- \* [Training with Prompts](../../examples/training/prompts/README.html)
- \* [What are Prompts?](../../examples/training/prompts/README.html#what-are-prompts)
  - \* [Why would we train with Prompts?](../../examples/training/prompts/README.html#why-would-we-train-with-prompts)
  - \* [How do we train with Prompts?](../../examples/training/prompts/README.html#how-do-we-train-with-prompts)
- \* [Training with PEFT Adapters](../../examples/training/peft/README.html)
- \* [Compatibility Methods](../../examples/training/peft/README.html#compatibility-methods)
- \* [Adding a New Adapter](../../examples/training/peft/README.html#adding-a-new-adapter)
  - \* [Loading a Pretrained Adapter](../../examples/training/peft/README.html#loading-a-pretrained-adapter)
- \* [Training Script](../../examples/training/peft/README.html#training-script)
- \* [Unsupervised Learning](../../examples/unsupervised\_learning/README.html)
- \* [TSDAE](../../examples/unsupervised\_learning/README.html#tsdae)
- \* [SimCSE](../../examples/unsupervised\_learning/README.html#simcse)
- \* [CT](../../examples/unsupervised\_learning/README.html#ct)
  - \* [CT (In-Batch Negative Sampling)](../../examples/unsupervised\_learning/README.html#ct-in-batch-negative-sampling)
  - \* [Masked Language Model (MLM)](../../examples/unsupervised\_learning/README.html#masked-language-model-mlm)
- \* [GenQ](../../examples/unsupervised\_learning/README.html#genq)
- \* [GPL](../../examples/unsupervised\_learning/README.html#gpl)
  - \* [Performance

[Comparison\]\(../../examples/unsupervised\\_learning/README.html#performance-comparison\)](#)

\* [\[Domain Adaptation\]\(../../examples/domain\\_adaptation/README.html\)](#)

\* [\[Domain Adaptation vs. Unsupervised Learning\]\(../../examples/domain\\_adaptation/README.html#domain-adaptation-vs-unsupervised-learning\)](#)

\* [\[Adaptive Pre-Training\]\(../../examples/domain\\_adaptation/README.html#adaptive-pre-training\)](#)

\* [\[GPL: Generative Pseudo-Labeling\]\(../../examples/domain\\_adaptation/README.html#gpl-generative-pseudo-labeling\)](#)

\* [\[Hyperparameter Optimization\]\(../../examples/training/hpo/README.html\)](#)

\* [\[HPO Components\]\(../../examples/training/hpo/README.html#hpo-components\)](#)

\* [\[Putting It All Together\]\(../../examples/training/hpo/README.html#putting-it-all-together\)](#)

\* [\[Example Scripts\]\(../../examples/training/hpo/README.html#example-scripts\)](#)

\* [\[Distributed Training\]\(../../sentence\\_transformer/training/distributed.html\)](#)

\* [\[Comparison\]\(../../sentence\\_transformer/training/distributed.html#comparison\)](#)

\* [\[FSDP\]\(../../sentence\\_transformer/training/distributed.html#fsdp\)](#)

## Cross Encoder

\* [\[Usage\]\(../../cross\\_encoder/usage/usage.html\)](#)

\* [\[Retrieve & Re-Rank\]\(../../examples/applications/retrieve\\_rerank/README.html\)](#)

\* [\[Retrieve & Re-Rank Pipeline\]\(../../examples/applications/retrieve\\_rerank/README.html#retrieve-re-rank-pipeline\)](#)

\* [\[Retrieval: Bi-Encoder\]\(../../examples/applications/retrieve\\_rerank/README.html#retrieval-bi-encoder\)](#)

\* [\[Re-Ranker:](#)

Cross-Encoder](../../examples/applications/retrieve\_rerank/README.html#re-ranker-cross-encoder)

\* [Example Scripts](../../examples/applications/retrieve\_rerank/README.html#example-scripts)

\* [Pre-trained Bi-Encoders (Retrieval)](../../examples/applications/retrieve\_rerank/README.html#pre-trained-bi-encoders-retrieval)

\* [Pre-trained Cross-Encoders (Re-Ranker)](../../examples/applications/retrieve\_rerank/README.html#pre-trained-cross-encoders-re-ranker)

\* [Pretrained Models](../../cross\_encoder/pretrained\_models.html)

\* [MS MARCO](../../cross\_encoder/pretrained\_models.html#ms-marco)

\* [SQuAD (QNLI)](../../cross\_encoder/pretrained\_models.html#squad-qnli)

\* [STSbenchmark](../../cross\_encoder/pretrained\_models.html#stsbenchmark)

\* [Quora Duplicate Questions](../../cross\_encoder/pretrained\_models.html#quora-duplicate-questions)

\* [NLI](../../cross\_encoder/pretrained\_models.html#nli)

\* [Community Models](../../cross\_encoder/pretrained\_models.html#community-models)

\* [Training Overview](../../cross\_encoder/training\_overview.html)

\* [Training Examples](../../cross\_encoder/training/examples.html)

\* [MS MARCO](../../examples/training/ms\_marco/cross\_encoder\_README.html)

\*

[Cross-Encoder](../../examples/training/ms\_marco/cross\_encoder\_README.html#cross-encoder)

\* [Cross-Encoder Knowledge Distillation](../../examples/training/ms\_marco/cross\_encoder\_README.html#cross-encoder-knowledge-distillation)

Package Reference

\* [Sentence Transformer](index.html)

\* [SentenceTransformer](SentenceTransformer.html)

\* [SentenceTransformer](SentenceTransformer.html#id1)

\*

[SentenceTransformerModelCardData](SentenceTransformer.html#sentencetransformermodelcarddata)

\* [SimilarityFunction](SentenceTransformer.html#similarityfunction)

\* [Trainer](trainer.html)

\* [SentenceTransformerTrainer](trainer.html#sentencetransformertrainer)

\* Training Arguments

\* SentenceTransformerTrainingArguments

\* [Losses](losses.html)

\* [BatchAllTripletLoss](losses.html#batchalltripletloss)

\* [BatchHardSoftMarginTripletLoss](losses.html#batchhardsoftmargintripletloss)

\* [BatchHardTripletLoss](losses.html#batchhardtripletloss)

\* [BatchSemiHardTripletLoss](losses.html#batchsemihardtripletloss)

\* [ContrastiveLoss](losses.html#contrastiveloss)

\* [OnlineContrastiveLoss](losses.html#onlinecontrastiveloss)

\* [ContrastiveTensionLoss](losses.html#contrastivetensionloss)

\*

[ContrastiveTensionLossInBatchNegatives](losses.html#contrastivetensionlossinbatchnegatives)

\* [CoSENTLoss](losses.html#cosentloss)

\* [AngleELoss](losses.html#angleloss)

\* [CosineSimilarityLoss](losses.html#cosinesimilarityloss)

\* [DenoisingAutoEncoderLoss](losses.html#denoisingautoencoderloss)

\* [GISTEmbedLoss](losses.html#gistembedloss)

- \* [CachedGISTEmbedLoss](losses.html#cachedgistembedloss)
- \* [MSELoss](losses.html#mseloss)
- \* [MarginMSELoss](losses.html#marginmseloss)
- \* [MatryoshkaLoss](losses.html#matryoshkaloss)
- \* [Matryoshka2dLoss](losses.html#matryoshka2dloss)
- \* [AdaptiveLayerLoss](losses.html#adaptivelayerloss)
- \* [MegaBatchMarginLoss](losses.html#megabatchmarginloss)
- \* [MultipleNegativesRankingLoss](losses.html#multiplenegativesrankingloss)
- \* [CachedMultipleNegativesRankingLoss](losses.html#cachedmultiplenegativesrankingloss)

\*

[MultipleNegativesSymmetricRankingLoss](losses.html#multiplenegativessymmetricrankingloss)

\*

[CachedMultipleNegativesSymmetricRankingLoss](losses.html#cachedmultiplenegativessymmetricrankingloss)

- \* [SoftmaxLoss](losses.html#softmaxloss)
- \* [TripletLoss](losses.html#tripletloss)
- \* [Samplers](sampler.html)
- \* [BatchSamplers](sampler.html#batchsamplers)
- \* [MultiDatasetBatchSamplers](sampler.html#multidatasetbatchsamplers)
- \* [Evaluation](evaluation.html)
- \* [BinaryClassificationEvaluator](evaluation.html#binaryclassificationevaluator)
- \* [EmbeddingSimilarityEvaluator](evaluation.html#embeddingsimilarityevaluator)
- \* [InformationRetrievalEvaluator](evaluation.html#informationretrievalevaluator)
- \* [NanoBEIREvaluator](evaluation.html#nanobeirevaluator)
- \* [MSEEvaluator](evaluation.html#mseevaluator)
- \* [ParaphraseMiningEvaluator](evaluation.html#paraphraseminingevaluator)
- \* [RerankingEvaluator](evaluation.html#rerankingevaluator)

- \* [SentenceEvaluator](evaluation.html#sentenceevaluator)
- \* [SequentialEvaluator](evaluation.html#sequentialevaluator)
- \* [TranslationEvaluator](evaluation.html#translationevaluator)
- \* [TripletEvaluator](evaluation.html#tripletevaluator)
- \* [Datasets](datasets.html)
- \* [ParallelSentencesDataset](datasets.html#parallelsentencesdataset)
- \* [SentenceLabelDataset](datasets.html#sentencelabeldataset)
- \* [DenoisingAutoEncoderDataset](datasets.html#denoisingautoencoderdataset)
- \* [NoDuplicatesDataLoader](datasets.html#noduplicatesdataloader)
- \* [Models](models.html)
- \* [Main Classes](models.html#main-classes)
- \* [Further Classes](models.html#further-classes)
- \* [quantization](quantization.html)

\*

[`quantize\_embeddings()`](quantization.html#sentence\_transformers.quantization.quantize\_embeddings)

\*

[`semantic\_search\_faiss()`](quantization.html#sentence\_transformers.quantization.semantic\_search\_faiss)

\*

[`semantic\_search\_usearch()`](quantization.html#sentence\_transformers.quantization.semantic\_search\_usearch)

- \* [Cross Encoder](../cross\_encoder/index.html)
- \* [CrossEncoder](../cross\_encoder/cross\_encoder.html)
- \* [CrossEncoder](../cross\_encoder/cross\_encoder.html#id1)
- \* [Training Inputs](../cross\_encoder/cross\_encoder.html#training-inputs)
- \* [Evaluation](../cross\_encoder/evaluation.html)

\* [CEBinaryAccuracyEvaluator](../cross\_encoder/evaluation.html#cebinaryaccuracyevaluator)

\*

[CEBinaryClassificationEvaluator](../cross\_encoder/evaluation.html#cebinaryclassificationevaluator)

\* [CECorrelationEvaluator](../cross\_encoder/evaluation.html#cecorrelationevaluator)

\* [CEF1Evaluator](../cross\_encoder/evaluation.html#cef1evaluator)

\*

[CESoftmaxAccuracyEvaluator](../cross\_encoder/evaluation.html#cesoftmaxaccuracyevaluator)

\* [CERerankingEvaluator](../cross\_encoder/evaluation.html#cererankingevaluator)

\* [util](../util.html)

\* [Helper Functions](../util.html#module-sentence\_transformers.util)

\* [community\_detection()](../util.html#sentence\_transformers.util.community\_detection)

\* [http\_get()](../util.html#sentence\_transformers.util.http\_get)

\* [is\_training\_available()](../util.html#sentence\_transformers.util.is\_training\_available)

\* [mine\_hard\_negatives()](../util.html#sentence\_transformers.util.mine\_hard\_negatives)

\* [normalize\_embeddings()](../util.html#sentence\_transformers.util.normalize\_embeddings)

\* [paraphrase\_mining()](../util.html#sentence\_transformers.util.paraphrase\_mining)

\* [semantic\_search()](../util.html#sentence\_transformers.util.semantic\_search)

\* [truncate\_embeddings()](../util.html#sentence\_transformers.util.truncate\_embeddings)

\* [Model Optimization](../util.html#module-sentence\_transformers.backend)

\*

[export\_dynamic\_quantized\_onnx\_model()](../util.html#sentence\_transformers.backend.export\_dynamic\_quantized\_onnx\_model)

\*

[export\_optimized\_onnx\_model()](../util.html#sentence\_transformers.backend.export\_optimized\_onnx\_model)

\*

[export\_static\_quantized\_openvino\_model()](../util.html#sentence\_transformers.backend.export\_st

atic\_quantized\_openvino\_model)

- \* [Similarity Metrics](../util.html#module-sentence\_transformers.util)
- \* [`cos_sim()`](../util.html#sentence\_transformers.util.cos\_sim)
- \* [`dot_score()`](../util.html#sentence\_transformers.util.dot\_score)
- \* [`euclidean_sim()`](../util.html#sentence\_transformers.util.euclidean\_sim)
- \* [`manhattan_sim()`](../util.html#sentence\_transformers.util.manhattan\_sim)
- \* [`pairwise_cos_sim()`](../util.html#sentence\_transformers.util.pairwise\_cos\_sim)
- \* [`pairwise_dot_score()`](../util.html#sentence\_transformers.util.pairwise\_dot\_score)
- \* [`pairwise_euclidean_sim()`](../util.html#sentence\_transformers.util.pairwise\_euclidean\_sim)
- \* [`pairwise_manhattan_sim()`](../util.html#sentence\_transformers.util.pairwise\_manhattan\_sim)

\_\_[Sentence Transformers](../../index.html)

\* [(../../index.html)]

\* [Sentence Transformer](index.html)

\* Training Arguments

\* [ [Edit on](#)

GitHub](https://github.com/UKPLab/sentence-transformers/blob/master/docs/package\_reference/sentence\_transformer/training\_args.md)

\* \* \*

# Training Arguments¶

## SentenceTransformerTrainingArguments¶

\_class \_sentence\_transformers.training\_args.SentenceTransformerTrainingArguments(\_output\_dir:



str\_, \_overwrite\_output\_dir: bool = False\_, \_do\_train: bool = False\_, \_do\_eval: bool = False\_,  
\_do\_predict: bool = False\_, \_eval\_strategy: ~transformers.trainer\_utils.IntervalStrategy | str = 'no',  
\_prediction\_loss\_only: bool = False\_, \_per\_device\_train\_batch\_size: int = 8\_,  
\_per\_device\_eval\_batch\_size: int = 8\_, \_per\_gpu\_train\_batch\_size: int | None = None\_,  
\_per\_gpu\_eval\_batch\_size: int | None = None\_, \_gradient\_accumulation\_steps: int = 1\_,  
\_eval\_accumulation\_steps: int | None = None\_, \_eval\_delay: float | None = 0\_,  
\_torch\_empty\_cache\_steps: int | None = None\_, \_learning\_rate: float = 5e-05\_, \_weight\_decay:  
float = 0.0\_, \_adam\_beta1: float = 0.9\_, \_adam\_beta2: float = 0.999\_, \_adam\_epsilon: float =  
1e-08\_, \_max\_grad\_norm: float = 1.0\_, \_num\_train\_epochs: float = 3.0\_, \_max\_steps: int = -1\_,  
\_lr\_scheduler\_type: ~transformers.trainer\_utils.SchedulerType | str = 'linear',  
\_lr\_scheduler\_kwargs: dict | str | None = <factory>\_, \_warmup\_ratio: float = 0.0\_, \_warmup\_steps:  
int = 0\_, \_log\_level: str | None = 'passive', \_log\_level\_replica: str | None = 'warning',  
\_log\_on\_each\_node: bool = True\_, \_logging\_dir: str | None = None\_, \_logging\_strategy:  
~transformers.trainer\_utils.IntervalStrategy | str = 'steps', \_logging\_first\_step: bool = False\_,  
\_logging\_steps: float = 500\_, \_logging\_nan\_inf\_filter: bool = True\_, \_save\_strategy:  
~transformers.trainer\_utils.IntervalStrategy | str = 'steps', \_save\_steps: float = 500\_,  
\_save\_total\_limit: int | None = None\_, \_save\_safetensors: bool | None = True\_,  
\_save\_on\_each\_node: bool = False\_, \_save\_only\_model: bool = False\_,  
\_restore\_callback\_states\_from\_checkpoint: bool = False\_, \_no\_cuda: bool = False\_, \_use\_cpu:  
bool = False\_, \_use\_mps\_device: bool = False\_, \_seed: int = 42\_, \_data\_seed: int | None = None\_,  
\_jit\_mode\_eval: bool = False\_, \_use\_ipex: bool = False\_, \_bf16: bool = False\_, \_fp16: bool =  
False\_, \_fp16\_opt\_level: str = 'O1', \_half\_precision\_backend: str = 'auto', \_bf16\_full\_eval: bool =  
False\_, \_fp16\_full\_eval: bool = False\_, \_tf32: bool | None = None\_, \_local\_rank: int = -1\_,  
\_ddp\_backend: str | None = None\_, \_tpu\_num\_cores: int | None = None\_, \_tpu\_metrics\_debug:  
bool = False\_, \_debug: str | ~typing.List[~transformers.debug\_utils.DebugOption] = "",  
\_dataloader\_drop\_last: bool = False\_, \_eval\_steps: float | None = None\_,  
\_dataloader\_num\_workers: int = 0\_, \_dataloader\_prefetch\_factor: int | None = None\_, \_past\_index:

int = -1\_, \_run\_name: str | None = None\_, \_disable\_tqdm: bool | None = None\_,  
\_remove\_unused\_columns: bool | None = True\_, \_label\_names: ~typing.List[str] | None = None\_,  
\_load\_best\_model\_at\_end: bool | None = False\_, \_metric\_for\_best\_model: str | None = None\_,  
\_greater\_is\_better: bool | None = None\_, \_ignore\_data\_skip: bool = False\_, \_fsdp:  
~typing.List[~transformers.trainer\_utils.FSDPOption] | str | None = "", \_fsdp\_min\_num\_params: int  
= 0\_, \_fsdp\_config: dict | str | None = None\_, \_fsdp\_transformer\_layer\_cls\_to\_wrap: str | None =  
None\_, \_accelerator\_config: dict | str | None = None\_, \_deepspeed: dict | str | None = None\_,  
\_label\_smoothing\_factor: float = 0.0\_, \_optim: ~transformers.training\_args.OptimizerNames | str =  
'adamw\_torch', \_optim\_args: str | None = None\_, \_adafactor: bool = False\_, \_group\_by\_length:  
bool = False\_, \_length\_column\_name: str | None = 'length', \_report\_to: None | str | ~typing.List[str]  
= None\_, \_ddp\_find\_unused\_parameters: bool | None = None\_, \_ddp\_bucket\_cap\_mb: int | None =  
None\_, \_ddp\_broadcast\_buffers: bool | None = None\_, \_dataloader\_pin\_memory: bool = True\_,  
\_dataloader\_persistent\_workers: bool = False\_, \_skip\_memory\_metrics: bool = True\_,  
\_use\_legacy\_prediction\_loop: bool = False\_, \_push\_to\_hub: bool = False\_,  
\_resume\_from\_checkpoint: str | None = None\_, \_hub\_model\_id: str | None = None\_, \_hub\_strategy:  
~transformers.trainer\_utils.HubStrategy | str = 'every\_save', \_hub\_token: str | None = None\_,  
\_hub\_private\_repo: bool = False\_, \_hub\_always\_push: bool = False\_, \_gradient\_checkpointing:  
bool = False\_, \_gradient\_checkpointing\_kwargs: dict | str | None = None\_,  
\_include\_inputs\_for\_metrics: bool = False\_, \_eval\_do\_concat\_batches: bool = True\_,  
\_fp16\_backend: str = 'auto', \_evaluation\_strategy: ~transformers.trainer\_utils.IntervalStrategy | str |  
None = None\_, \_push\_to\_hub\_model\_id: str | None = None\_, \_push\_to\_hub\_organization: str |  
None = None\_, \_push\_to\_hub\_token: str | None = None\_, \_mp\_parameters: str = "",  
\_auto\_find\_batch\_size: bool = False\_, \_full\_determinism: bool = False\_, \_torchdynamo: str | None =  
None\_, \_ray\_scope: str | None = 'last', \_ddp\_timeout: int | None = 1800\_, \_torch\_compile: bool =  
False\_, \_torch\_compile\_backend: str | None = None\_, \_torch\_compile\_mode: str | None = None\_,  
\_dispatch\_batches: bool | None = None\_, \_split\_batches: bool | None = None\_,  
\_include\_tokens\_per\_second: bool | None = False\_, \_include\_num\_input\_tokens\_seen: bool | None

```

= False_, _neptune_noise_alpha: float | None = None_, _optim_target_modules: None | str |
~typing.List[str] = None_, _batch_eval_metrics: bool = False_, _eval_on_start: bool = False_,
_eval_use_gather_object: bool | None = False_, _prompts: str | None = None_, _batch_sampler:
~sentence_transformers.training_args.BatchSamplers | str = BatchSamplers.BATCH_SAMPLER_,
_multi_dataset_batch_sampler: ~sentence_transformers.training_args.MultiDatasetBatchSamplers |
str
=
MultiDatasetBatchSamplers.PROPORTIONAL_) [[source]](https://github.com/UKPLab/sentence-tran
sformers/blob/master/sentence_transformers\\training_args.py#L143-L220)ïf•

```

SentenceTransformerTrainingArguments extends

[`TrainingArguments`](https://huggingface.co/docs/transformers/main/en/main\_classes/trainer#transformers.TrainingArguments)

"(in transformers vmain)") with additional arguments specific to Sentence

Transformers. See

[`TrainingArguments`](https://huggingface.co/docs/transformers/main/en/main\_classes/trainer#transformers.TrainingArguments)

"(in transformers vmain)") for the complete list of available arguments.

Parameters:

**output\_dir** (str) â€” The output directory where the model checkpoints will be written.

**prompts** (Union[Dict[str, Dict[str, str]], Dict[str, str], str, \_optional\_) â€”

The prompts to use for each column in the training, evaluation and test

datasets. Four formats are accepted:

1. str: A single prompt to use for all columns in the datasets, regardless of whether the training/evaluation/test datasets are `[ datasets.Dataset ]`([https://huggingface.co/docs/datasets/main/en/package\\_reference/main\\_classes](https://huggingface.co/docs/datasets/main/en/package_reference/main_classes) `#datasets.Dataset` `"\(\in datasets vmain\)"`) or a `[ datasets.DatasetDict ]`([https://huggingface.co/docs/datasets/main/en/package\\_reference/main\\_classes](https://huggingface.co/docs/datasets/main/en/package_reference/main_classes) `#datasets.DatasetDict` `"\(\in datasets vmain\)"`).

2. Dict[str, str]: A dictionary mapping column names to prompts, regardless of whether the training/evaluation/test datasets are `[ datasets.Dataset ]`([https://huggingface.co/docs/datasets/main/en/package\\_reference/main\\_classes](https://huggingface.co/docs/datasets/main/en/package_reference/main_classes) `#datasets.Dataset` `"\(\in datasets vmain\)"`) or a `[ datasets.DatasetDict ]`([https://huggingface.co/docs/datasets/main/en/package\\_reference/main\\_classes](https://huggingface.co/docs/datasets/main/en/package_reference/main_classes) `#datasets.DatasetDict` `"\(\in datasets vmain\)"`).

3. Dict[str, str]: A dictionary mapping dataset names to prompts. This should only be used if your training/evaluation/test datasets are a `[ datasets.DatasetDict ]`([https://huggingface.co/docs/datasets/main/en/package\\_reference/main\\_classes](https://huggingface.co/docs/datasets/main/en/package_reference/main_classes) `#datasets.DatasetDict` `"\(\in datasets vmain\)"`) or a dictionary of `[ datasets.Dataset ]`([https://huggingface.co/docs/datasets/main/en/package\\_reference/main\\_classes](https://huggingface.co/docs/datasets/main/en/package_reference/main_classes) `#datasets.Dataset` `"\(\in datasets vmain\)"`).

4. Dict[str, Dict[str, str]]: A dictionary mapping dataset names to dictionaries mapping column names to prompts. This should only be used if your training/evaluation/test datasets are a `[ datasets.DatasetDict ]`([https://huggingface.co/docs/datasets/main/en/package\\_reference/main\\_classes](https://huggingface.co/docs/datasets/main/en/package_reference/main_classes)

sses#datasets.DatasetDict "\(\in datasets vmain\)") or a dictionary of  
[ datasets.Dataset`](https://huggingface.co/docs/datasets/main/en/package\_reference/main\_classes  
#datasets.Dataset "\(\in datasets vmain\)").

\*

**\*\*batch\_sampler\*\***

(Union[[ `BatchSamplers`](sampler.html#sentence\_transformers.training\_args.BatchSamplers  
"sentence\_transformers.training\_args.BatchSamplers"), str], \_optional\_) â€” The batch sampler to  
use. See [ `BatchSamplers`](sampler.html#sentence\_transformers.training\_args.BatchSamplers  
"sentence\_transformers.training\_args.BatchSamplers") for valid options. Defaults to  
`BatchSamplers.BATCH\_SAMPLER`.

\*

**\*\*multi\_dataset\_batch\_sampler\*\***

(Union[[ `MultiDatasetBatchSamplers`](sampler.html#sentence\_transformers.training\_args.MultiData  
setBatchSamplers "sentence\_transformers.training\_args.MultiDatasetBatchSamplers"), str],  
\_optional\_) â€” The multi-dataset batch sampler to use. See  
[ `MultiDatasetBatchSamplers`](sampler.html#sentence\_transformers.training\_args.MultiDatasetBatc  
hSamplers "sentence\_transformers.training\_args.MultiDatasetBatchSamplers") for valid options.  
Defaults to `MultiDatasetBatchSamplers.PROPORTIONAL`.

\_property \_ddp\_timeout\_delta \_: timedelta\_ïf•

The actual timeout for torch.distributed.init\_process\_group since it expects a  
timedelta variable.

\_property \_device \_:

[device](https://pytorch.org/docs/stable/tensor\_attributes.html#torch.device

"\n(in PyTorch v2.5\n)")\_if•

The device used by this process.

\_property \_eval\_batch\_size \_: int\_if•

The actual batch size for evaluation (may differ from per\_gpu\_eval\_batch\_size in distributed training).

get\_process\_log\_level()if•

Returns the log level to be used depending on whether this process is the main process of node 0, main process of node non-0, or a non-main process.

For the main process the log level defaults to the logging level set (logging.WARNING if you didnâ€™t do anything) unless overridden by log\_level argument.

For the replica processes the log level defaults to logging.WARNING unless overridden by log\_level\_replica argument.

The choice between the main and replica process settings is made according to the return value of `should_log`.

```
get_warmup_steps(_num_training_steps : int_) if •
```

Get number of steps used for a linear warmup.

```
_property _local_process_index if •
```

The index of the local process used.

```
main_process_first(_local = True, _desc = 'work') if •
```

A context manager for torch distributed environment where one needs to do something on the main process, while blocking replicas, and when it's finished releasing the replicas.

One such use is for datasets' map feature which to be efficient should be run once on the main process, which upon completion saves a cached version of results and which then automatically gets loaded by the replicas.

Parameters:

\* **local** (bool, \_optional\_ , defaults to True) – if True first means process of rank 0 of each node if False first means process of rank 0 of node rank 0 In multi-node environment with a shared filesystem you most likely will want to use local=False so that only the main process of the first node will do the processing. If however, the filesystem is not shared, then the main process of each node will need to do the processing, which is the default behavior.

\* **desc** (str, \_optional\_ , defaults to 'œworkœ') – a work description to be used in debug logs

\_property \_n\_gpus•

The number of GPUs used by this process.

Note

This will only be greater than one when you have multiple GPUs available but are not using distributed training. For distributed training, it will always be 1.

\_property \_parallel\_mode•



The current mode used for parallelism if multiple GPUs/TPU cores are available. One of:

- \* `ParallelMode.NOT_PARALLEL`: no parallelism (CPU or one GPU).

- \* `ParallelMode.NOT_DISTRIBUTED`: several GPUs in one single process (uses `torch.nn.DataParallel`).

- \* `ParallelMode.DISTRIBUTED`: several GPUs, each having its own process (uses `torch.nn.DistributedDataParallel`).

- \* `ParallelMode.TPU`: several TPU cores.

`_property _place_model_on_device`•

Can be subclassed and overridden for some specific integrations.

`_property _process_index`•

The index of the current process used.

```
set_dataloader(_train_batch_size : int = 8_, _eval_batch_size : int = 8_, _drop_last : bool = False_,
               _num_workers : int = 0_, _pin_memory : bool = True_, _persistent_workers : bool = False_,
               _prefetch_factor : int | None = None_, _auto_find_batch_size : bool = False_, _ignore_data_skip :
               bool = False_, _sampler_seed : int | None = None_)if•
```

A method that regroups all arguments linked to the dataloaders creation.

Parameters:

\* **drop\_last** (bool, \_optional\_ , defaults to False) â€œ Whether to drop the last incomplete batch (if the length of the dataset is not divisible by the batch size) or not.

\* **num\_workers** (int, \_optional\_ , defaults to 0) â€œ Number of subprocesses to use for data loading (PyTorch only). 0 means that the data will be loaded in the main process.

\* **pin\_memory** (bool, \_optional\_ , defaults to True) â€œ Whether you want to pin memory in data loaders or not. Will default to True.

\* **persistent\_workers** (bool, \_optional\_ , defaults to False) â€œ If True, the data loader will not shut down the worker processes after a dataset has been consumed once. This allows to maintain the workers Dataset instances alive. Can potentially speed up training, but will increase RAM usage. Will default to False.

\* **prefetch\_factor** (int, \_optional\_) â€œ Number of batches loaded in advance by each worker. 2 means there will be a total of 2 \* num\_workers batches prefetched across all workers.

\* **auto\_find\_batch\_size** (bool, \_optional\_ , defaults to False) â€œ Whether to find a batch size that will fit into memory automatically through exponential decay, avoiding CUDA Out-of-Memory errors. Requires accelerate to be installed (pip install accelerate)

\* **ignore\_data\_skip** (bool, \_optional\_ , defaults to False) â€œ When resuming training, whether or not to skip the epochs and batches to get the data loading at the same stage as in the previous training. If set to True, the training will begin faster (as that skipping step can take a long time) but will not yield the same results as the interrupted training would have.

\* **sampler\_seed** (int, \_optional\_) â€œ Random seed to be used with data samplers. If not set, random generators for data sampling will use the same seed as self.seed. This can be used to ensure reproducibility of data sampling, independent of the model seed.

Example:

```
```py >>> from transformers import TrainingArguments
```

```
>>> args = TrainingArguments("working_dir")
```

```
>>> args = args.set_dataloader(train_batch_size=16, eval_batch_size=64)
```

```
>>> args.per_device_train_batch_size
```

```
16
```

```
...
```

```

set_evaluate(_strategy : str | IntervalStrategy = 'no', _steps : int = 500, _batch_size : int = 8,
_accumulation_steps : int | None = None, _delay : float | None = None, _loss_only : bool =
False, _jit_mode : bool = False)if

```

A method that regroups all arguments linked to evaluation.

Parameters:

```

* **strategy** (str or [~trainer_utils.IntervalStrategy], _optional_ , defaults to "no")

```

The evaluation strategy to adopt during training. Possible values are:

- > \* "no": No evaluation is done during training.
- >
- > \* "steps": Evaluation is done (and logged) every steps.
- >
- > \* "epoch": Evaluation is done at the end of each epoch.

Setting a strategy different from "no" will set self.do\_eval to True.

```

* **steps** (int, _optional_ , defaults to 500)

```

Number of update steps between two evaluations if strategy="steps".

\* **batch\_size** (int, \_optional\_) â€” The batch size per device (GPU/TPU core/CPU) used for evaluation.

\* **accumulation\_steps** (int, \_optional\_) â€” Number of predictions steps to accumulate the output tensors for, before moving the results to the CPU. If left unset, the whole predictions are accumulated on GPU/TPU before being moved to the CPU (faster but requires more memory).

\* **delay** (float, \_optional\_) â€” Number of epochs or steps to wait for before the first evaluation can be performed, depending on the eval\_strategy.

\* **loss\_only** (bool, \_optional\_, defaults to False) â€” Ignores all outputs except the loss.

\* **jit\_mode** (bool, \_optional\_) â€” Whether or not to use PyTorch jit trace for inference.

Example:

```
```py >>> from transformers import TrainingArguments
```

```
>>> args = TrainingArguments("working_dir")
```

```
>>> args = args.set_evaluate(strategy="steps", steps=100)
```

```
>>> args.eval_steps
```

```
100
```

```
```
```

```
set_logging(_strategy : str | IntervalStrategy = 'steps', _steps : int = 500, _report_to : str | List[str]
= 'none', _level : str = 'passive', _first_step : bool = False, _nan_inf_filter : bool = False,
_on_each_node : bool = False, _replica_level : str = 'passive')if•
```

A method that regroups all arguments linked to logging.

Parameters:

**strategy** (str or [~trainer\_utils.IntervalStrategy], \_optional\_ , defaults to "steps")

The logging strategy to adopt during training. Possible values are:

- > \* "no": No logging is done during training.
- >
- > \* "epoch": Logging is done at the end of each epoch.
- >
- > \* "steps": Logging is done every logging\_steps.

**steps** (int, \_optional\_ , defaults to 500) " Number of update steps between two logs if strategy="steps".

**level** (str, \_optional\_ , defaults to "passive") " Logger log level to use on the main process. Possible choices are the log levels as strings: "debug", "info", "warning",

â€œerrorâ€• and â€œcriticalâ€•, plus a â€œpassiveâ€• level which doesnâ€™t set anything and lets the application set the level.

\* **report\_to** (str or List[str], \_optional\_ , defaults to â€œallâ€•) â€œ The list of integrations to report the results and logs to. Supported platforms are â€œazure\_mlâ€•, â€œclearmlâ€•, â€œcodecarbonâ€•, â€œcomet\_mlâ€•, â€œdagshubâ€•, â€œdvcliveâ€•, â€œflyteâ€•, â€œmlflowâ€•, â€œneptuneâ€•, â€œtensorboardâ€•, and â€œwandbâ€•. Use â€œallâ€• to report to all integrations installed, â€œnoneâ€• for no integrations.

\* **first\_step** (bool, \_optional\_ , defaults to False) â€œ Whether to log and evaluate the first global\_step or not.

\* **nan\_inf\_filter** (bool, \_optional\_ , defaults to True) â€œ

Whether to filter nan and inf losses for logging. If set to True the loss of every step that is nan or inf is filtered and the average loss of the current logging window is taken instead.

<Tip>

nan\_inf\_filter only influences the logging of loss values, it does not change the behavior the gradient is computed or applied to the model.

</Tip>

\* **on\_each\_node** (bool, \_optional\_ , defaults to True) â€œ In multinode distributed training, whether to log using log\_level once per node, or only on the main node.

\* **replica\_level** (str, \_optional\_ , defaults to "passive") " Logger log level to use on replicas. Same choices as log\_level

Example:

```
```py >>> from transformers import TrainingArguments
```

```
>>> args = TrainingArguments("working_dir")
```

```
>>> args = args.set_logging(strategy="steps", steps=100)
```

```
>>> args.logging_steps
```

```
100
```

```
```
```

set\_lr\_scheduler(\_name : str | SchedulerType = 'linear', \_num\_epochs : float = 3.0, \_max\_steps : int = -1, \_warmup\_ratio : float = 0, \_warmup\_steps : int = 0)if•

A method that regroups all arguments linked to the learning rate scheduler and its hyperparameters.

Parameters:



\* **name** (str or [SchedulerType], \_optional\_ , defaults to "linear") â€” The scheduler type to use. See the documentation of [SchedulerType] for all possible values.

\* **num\_epochs** (float, \_optional\_ , defaults to 3.0) â€” Total number of training epochs to perform (if not an integer, will perform the decimal part percents of the last epoch before stopping training).

\* **max\_steps** (int, \_optional\_ , defaults to -1) â€” If set to a positive number, the total number of training steps to perform. Overrides num\_train\_epochs. For a finite dataset, training is reiterated through the dataset (if all data is exhausted) until max\_steps is reached.

\* **warmup\_ratio** (float, \_optional\_ , defaults to 0.0) â€” Ratio of total training steps used for a linear warmup from 0 to learning\_rate.

\* **warmup\_steps** (int, \_optional\_ , defaults to 0) â€” Number of steps used for a linear warmup from 0 to learning\_rate. Overrides any effect of warmup\_ratio.

Example:

```
```py >>> from transformers import TrainingArguments
```

```
>>> args = TrainingArguments("working_dir")
```

```
>>> args = args.set_lr_scheduler(name="cosine", warmup_ratio=0.05)
```

```
>>> args.warmup_ratio
```

0.05

...

```
set_optimizer(_name : str | OptimizerNames = 'adamw_torch', _learning_rate : float = 5e-05_,  
_weight_decay : float = 0_, _beta1 : float = 0.9_, _beta2 : float = 0.999_, _epsilon : float = 1e-08_,  
_args : str | None = None_)if•
```

A method that regroups all arguments linked to the optimizer and its hyperparameters.

Parameters:

\* **name** (str or [training\_args.OptimizerNames], \_optional\_ , defaults to "adamw\_torch")  
" The optimizer to use: "adamw\_hf", "adamw\_torch", "adamw\_torch\_fused",  
"adamw\_apex\_fused", "adamw\_anyprecision" or "adafactor".

\* **learning\_rate** (float, \_optional\_ , defaults to 5e-5) " The initial learning rate.

\* **weight\_decay** (float, \_optional\_ , defaults to 0) " The weight decay to apply (if not zero) to all layers except all bias and LayerNorm weights.

\* **beta1** (float, \_optional\_ , defaults to 0.9) " The beta1 hyperparameter for the adam optimizer or its variants.

\* **beta2** (float, \_optional\_ , defaults to 0.999) â€” The beta2 hyperparameter for the adam optimizer or its variants.

\* **epsilon** (float, \_optional\_ , defaults to 1e-8) â€” The epsilon hyperparameter for the adam optimizer or its variants.

\* **args** (str, \_optional\_) â€” Optional arguments that are supplied to AnyPrecisionAdamW (only useful when optim=â€•adamw\_anyprecisionâ€•).

Example:

```
``py >>> from transformers import TrainingArguments
```

```
>>> args = TrainingArguments("working_dir")
```

```
>>> args = args.set_optimizer(name="adamw_torch", beta1=0.8)
```

```
>>> args.optim
```

```
'adamw_torch'
```

```
...
```

```
set_push_to_hub(_model_id : str, _strategy : str | HubStrategy = 'every_save', _token : str | None = None, _private_repo : bool = False, _always_push : bool = False)if•
```

A method that regroups all arguments linked to synchronizing checkpoints with the Hub.

<Tip>

Calling this method will set `self.push_to_hub` to `True`, which means the `output_dir` will begin a git directory synced with the repo (determined by `model_id`) and the content will be pushed each time a save is triggered (depending on your `self.save_strategy`). Calling `[~Trainer.save_model]` will also trigger a push.

</Tip>

Parameters:

\* **`model_id`** (str) â€œ The name of the repository to keep in sync with the local `_output_dir_`. It can be a simple model ID in which case the model will be pushed in your namespace. Otherwise it should be the whole repository name, for instance `â€œuser_name/modelâ€•`, which allows you to push to an organization you are a member of with `â€œorganization_name/modelâ€•`.

\* **`strategy`** (str or `[~trainer_utils.HubStrategy]`, `_optional_`, defaults to `â€œevery_saveâ€•`) â€œ

Defines the scope of what is pushed to the Hub and when. Possible values are:

\* `â€•endâ€•`: push the model, its configuration, the tokenizer (if passed along to the `[Trainer]`) and

a

draft of a model card when the [~Trainer.save\_model] method is called. \-

â€œevery\_saveâ€•: push the model, its configuration, the tokenizer (if passed along to the [Trainer])

> and

a draft of a model card each time there is a model save. The pushes are asynchronous to not block training, and in case the save are very frequent, a new push is only attempted if the previous one is finished. A last push is made with the final model at the end of training. \- â€œcheckpointâ€•: like â€œevery\_saveâ€• but the latest checkpoint is also pushed in a subfolder named last-checkpoint, allowing you to resume training easily with `trainer.train(resume_from_checkpoint=â€•last-checkpointâ€•)`. \- â€œall\_checkpointsâ€•: like â€œcheckpointâ€• but all checkpoints are pushed like they appear in the

> output

folder (so you will get one checkpoint folder per folder in your final repository)

\* **token** (str, \_optional\_) â€“ The token to use to push the model to the Hub. Will default to the token in the cache folder obtained with `huggingface-cli login`.

\* **private\_repo** (bool, \_optional\_ , defaults to False) â€“ If True, the Hub repo will be set to

private.

\* **always\_push** (bool, \_optional\_ , defaults to False) – Unless this is True, the Trainer will skip pushing a checkpoint when the previous push is not finished.

Example:

```
```py >>> from transformers import TrainingArguments
```

```
>>> args = TrainingArguments("working_dir")
```

```
>>> args = args.set_push_to_hub("me/awesome-model")
```

```
>>> args.hub_model_id
```

```
'me/awesome-model'
```

```
...
```

`set_save(_strategy : str | IntervalStrategy = 'steps', _steps : int = 500, _total_limit : int | None = None, _on_each_node : bool = False)`•

A method that regroups all arguments linked to checkpoint saving.

Parameters:

\* **strategy** (str or [`~trainer_utils.IntervalStrategy`], \_optional\_ , defaults to `steps`)

The checkpoint save strategy to adopt during training. Possible values are:

- > \* `no`: No save is done during training.
- >
- > \* `epoch`: Save is done at the end of each epoch.
- >
- > \* `steps`: Save is done every `save_steps`.

\* **steps** (int, \_optional\_ , defaults to 500) Number of updates steps before two checkpoint saves if `strategy=steps`.

\* **total\_limit** (int, \_optional\_) If a value is passed, will limit the total amount of checkpoints. Deletes the older checkpoints in `output_dir`.

\* **on\_each\_node** (bool, \_optional\_ , defaults to False)

When doing multi-node distributed training, whether to save models and checkpoints on each node, or only on the main one.

This should not be activated when the different nodes use the same storage as the files will be saved with the same names for each node.

Example:

```
```py >>> from transformers import TrainingArguments
```

```
>>> args = TrainingArguments("working_dir")
```

```
>>> args = args.set_save(strategy="steps", steps=100)
```

```
>>> args.save_steps
```

```
100
```

```
```
```

```
set_testing(_batch_size : int = 8_, _loss_only : bool = False_, _jit_mode :  
bool = False_)if•
```

A method that regroups all basic arguments linked to testing on a held-out dataset.

<Tip>

Calling this method will automatically set self.do\_predict to True.

</Tip>

Parameters:



\* **batch\_size** (int \_optional\_ , defaults to 8) â€“ The batch size per device (GPU/TPU core/CPUâ€¦) used for testing.

\* **loss\_only** (bool, \_optional\_ , defaults to False) â€“ Ignores all outputs except the loss.

\* **jit\_mode** (bool, \_optional\_) â€“ Whether or not to use PyTorch jit trace for inference.

Example:

```
```py >>> from transformers import TrainingArguments
```

```
>>> args = TrainingArguments("working_dir")
```

```
>>> args = args.set_testing(batch_size=32)
```

```
>>> args.per_device_eval_batch_size
```

```
32
```

```
...
```

```
set_training(_learning_rate : float = 5e-05_, _batch_size : int = 8_,
```

```
_weight_decay : float = 0_, _num_epochs : float = 3_, _max_steps : int = -1_,
```

```
_gradient_accumulation_steps : int = 1_, _seed : int = 42_,
```

```
_gradient_checkpointing : bool = False_)if•
```

A method that regroups all basic arguments linked to the training.

<Tip>

Calling this method will automatically set self.do\_train to True.

</Tip>

Parameters:

- \* **learning\_rate** (float, \_optional\_ , defaults to 5e-5) â€” The initial learning rate for the optimizer.
- \* **batch\_size** (int \_optional\_ , defaults to 8) â€” The batch size per device (GPU/TPU core/CPUâ€¦) used for training.
- \* **weight\_decay** (float, \_optional\_ , defaults to 0) â€” The weight decay to apply (if not zero) to all layers except all bias and LayerNorm weights in the optimizer.
- \* **num\_train\_epochs** (float, \_optional\_ , defaults to 3.0) â€” Total number of training epochs to perform (if not an integer, will perform the decimal part percents of the last epoch before stopping training).
- \* **max\_steps** (int, \_optional\_ , defaults to -1) â€” If set to a positive number, the total number of

training steps to perform. Overrides num\_train\_epochs. For a finite dataset, training is reiterated through the dataset (if all data is exhausted) until max\_steps is reached.

**gradient\_accumulation\_steps** (int, \_optional\_ , defaults to 1) â€“

Number of updates steps to accumulate the gradients for, before performing a backward/update pass.

<Tip warning={true}>

When using gradient accumulation, one step is counted as one step with backward pass. Therefore, logging, evaluation, save will be conducted every gradient\_accumulation\_steps \* xxx\_step training examples.

</Tip>

**seed** (int, \_optional\_ , defaults to 42) â€“ Random seed that will be set at the beginning of training. To ensure reproducibility across runs, use the [~Trainer.model\_init] function to instantiate the model if it has some randomly initialized parameters.

**gradient\_checkpointing** (bool, \_optional\_ , defaults to False) â€“ If True, use gradient checkpointing to save memory at the expense of slower backward pass.

Example:

```
```py >>> from transformers import TrainingArguments
```

```
>>> args = TrainingArguments("working_dir")
>>> args = args.set_training(learning_rate=1e-4, batch_size=32)
>>> args.learning_rate
1e-4
...
```

`_property_should_log`•

Whether or not the current process should produce log.

`_property_should_save`•

Whether or not the current process should write to disk, e.g., to save models and checkpoints.

`to_dict()`•

Serializes this instance while replace Enum by their values (for JSON

serialization support). It obfuscates the token values by removing their value.

`to_json_string()`•

Serializes this instance to a JSON string.

`to_sanitized_dict()` -> Dict[str, Any]•

Sanitized serialization to use with TensorBoard™s hparams

`_property _train_batch_size _: int`•

The actual batch size for training (may differ from `per_gpu_train_batch_size` in distributed training).

`_property _world_size`•

The number of processes used in parallel.

[ Previous](trainer.html "Trainer") [Next ](losses.html "Losses")

\* \* \*

(C) Copyright 2025.

Built with [Sphinx](https://www.sphinx-doc.org/) using a  
[theme](https://github.com/readthedocs/sphinx\_rtd\_theme) provided by [Read the  
Docs](https://readthedocs.org).