

# Athena – Research Companion



# What is Athena?

AI-Assist to facilitate scientific Research with:

- Advanced Semantic Search
- Human-AI Collaboration
- Comprehensive Admin Support

# Cohere-powered Tasks

- Enrich Paper Abstracts w/ Knowledge-Base
- Prepare Glossary of Technical Terms
- Find similar Papers
- Find relevant Papers in a given Research Topic
- Compose E-mail to Authors
- Write a Tweet about a Paper

# Steps

- Dataset Creation
- Embeddings and Indexing w/ Weaviate
- Prompt Templates, Output-Formatting, Validate
- Cohere Engine Abstraction
- Demo App

# Datasets Creation w/ Embed-v3

- 50K ArXiv Papers' Metadata w/ Embeddings.
- Made available to the OpenSource community:
  - <https://huggingface.co/datasets/dcarpintero/arXiv.cs.AI.CL.CV.LG.MA.NE.embedv3>
  - <https://huggingface.co/datasets/dcarpintero/arXiv.cs.CL.embedv3>

# Enhanced Generation

- Prompt Templates
- Langchain Expression Language
- Pydantic

# https://athena-research.streamlit.app/

### Athena Research

ARXIV

Article ID

COHERE-SETTINGS

Generation Model

Embeddings Model

Rank Model

Max Results

WEAVIATE-SETTINGS

Cluster

[Github Repo](#) [Cohere LLMs](#) [Weaviate](#)

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding | Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova | 2019-05-24 | <http://arxiv.org/abs/1810.04805v2>

TL;DR [SIMILAR ARTICLES](#) [FINDER](#) [EMAIL-AUTHORS](#) [TWEET](#)

Lists the most similar Articles from a self-created [embeddings](#) arXiv dataset of 50k entries in AI, ML and NLP [indexed with Weaviate](#)

#### A text autoencoder from transformer for fast encoding language representation

In recent years BERT shows apparent advantages and great potential in natural language processing tasks. However, both training and applying BERT requires intensive time and resources for computing contextual language representations, which hinders its universality and applicability. To overcome this bottleneck, we propose a deep bidirectional language model by using window masking mechanism at attention layer. This work computes contextual language representations without random masking as does in BERT and maintains the deep bidirectional architecture like BERT. To compute the same sentence representation, our method shows  $O(n)$  complexity less compared to other transformer-based models with  $O(n^2)$ . To further demonstrate its superiority, computing context language representations on CPU [...] [\[2111.02844\]](#) [PDF](#)

#### Utilizing Bidirectional Encoder Representations from Transformers for Answer Selection

Pre-training a transformer-based model for the language modeling task in a large dataset and then fine-tuning it for downstream tasks has been found very useful in recent years. One major advantage of such pre-trained language models is that they can effectively absorb the context of each word in a sentence. However, for tasks such as the answer selection task, the pre-trained language models have not been extensively used yet. To investigate their effectiveness in such tasks, in this paper, we adopt the pre-trained Bidirectional Encoder Representations from Transformer (BERT) language model and fine-tune it on two Question Answering (QA) datasets and three Community Question Answering (CQA) datasets for the answer selection task. We find that fine-tuning the BERT model for the answer sele [...] [\[2011.07208\]](#) [PDF](#)

#### schuBERT: Optimizing Elements of BERT

Transformers [\citep{vaswani2017attention}](#) have gradually become a key component for many state-of-the-art natural language representation models. A recent Transformer based model- BERT [\citep{devlin2018bert}](#) achieved state-of-the-art results on various natural language processing tasks, including GLUE, SQuAD v1.1, and SQuAD v2.0. This model however is computationally prohibitive and has a huge number of parameters. In this work we revisit the architecture choices of BERT in efforts to obtain a lighter model. We focus on reducing the number of parameters yet our methods can be applied towards other objectives such FLOPs or latency. We show that much efficient light BERT models can be obtained by reducing algorithmically chosen correct architecture design dimensions rather than reducing the [...] [\[2005.06628\]](#) [PDF](#)

#### Mono vs Multilingual Transformer-based Models: a Comparison across Several Language Tasks

BERT (Bidirectional Encoder Representations from Transformers) and ALBERT (A Lite BERT) are methods for pre-training language models which can later be fine-tuned for a variety of Natural Language Understanding tasks. These methods have been applied to a number of such tasks (mostly in English), achieving results that outperform the state-of-the-art. In this paper, our contribution is twofold. First, we make available our trained BERT and Albert model for Portuguese. Second, we compare our monolingual and the standard multilingual models using experiments in semantic textual similarity, recognizing textual entailment, textual category classification, sentiment analysis, offensive comment detection, and fake news detection, to assess the effectiveness of the generated language representatio [...] [\[2007.09757\]](#) [PDF](#)