ME 5374-ST Fall 2025

Machine Learning for Materials Science and Discovery

Asst. Prof. Peter Schindler

Dr. Emad Rezaei

Lecture 16- Large Language Models

Northeastern University

Agenda

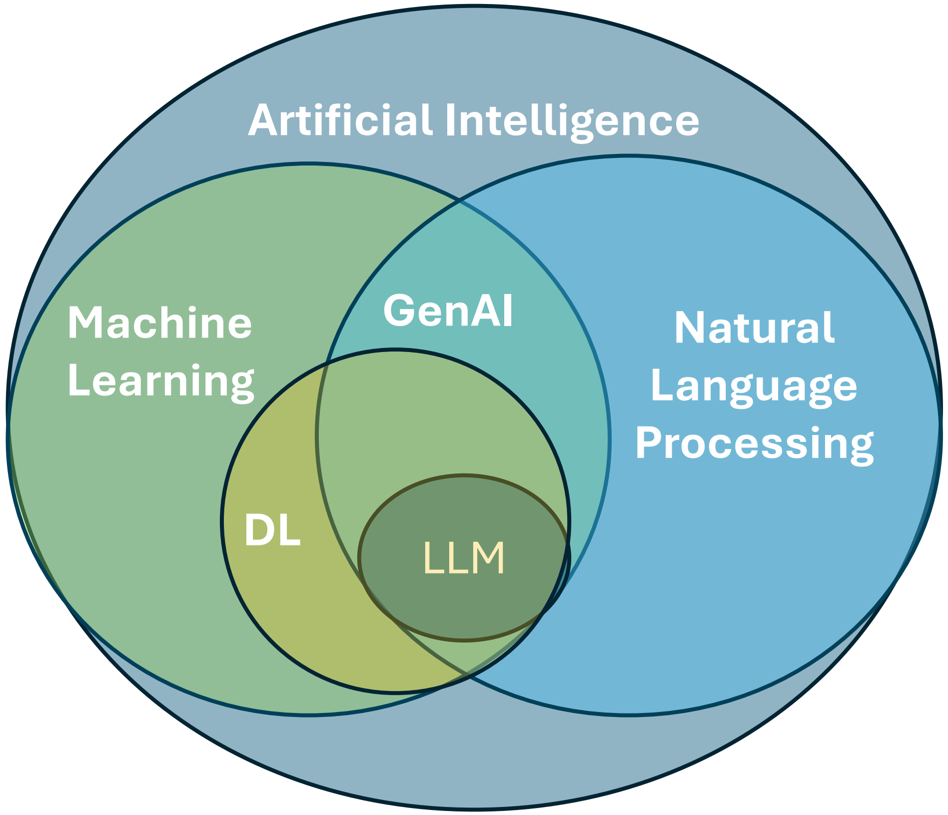
- Introduction and history of language models
- Methods and approaches
- Applications in Materials science

Language Models

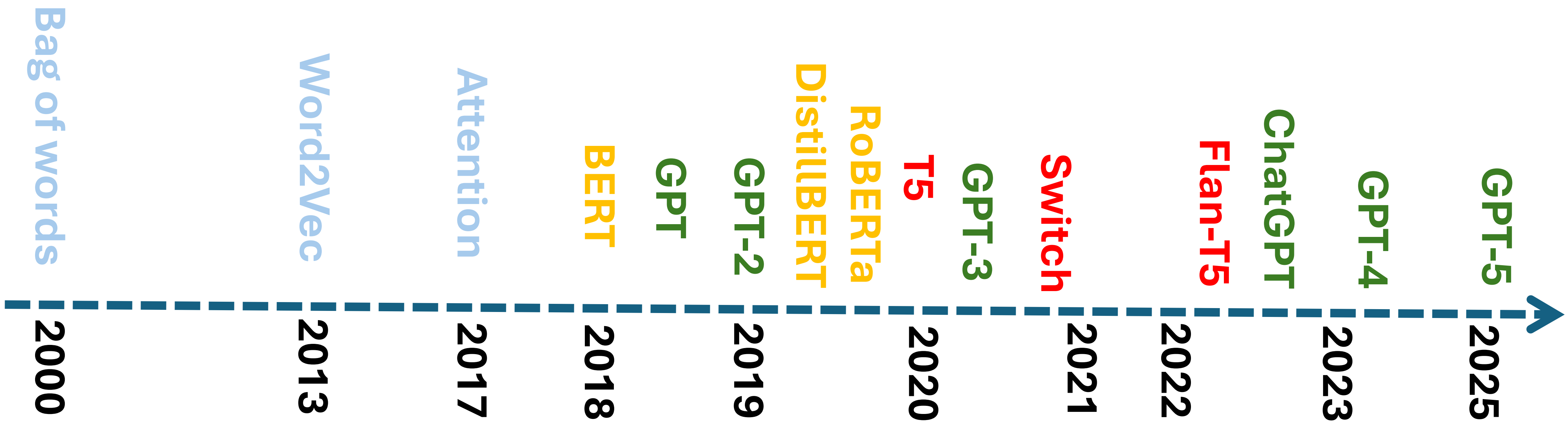
Language AI : a subfield of AI aimed to develop models that are able to understand, process, and generate human language.

Natural language processing (NLP): is a fundamental part of language AI.

NLP Focuses on specific tasks like text classification, sentiment analysis, but language AI performs a wider range of tasks e.g. language understanding, and content generation.



History of Language AI



Non-transformer

Decoder only

Encoder-only

Encoder-Decoder

How to represent language to computers

Text is unstructured! 😐

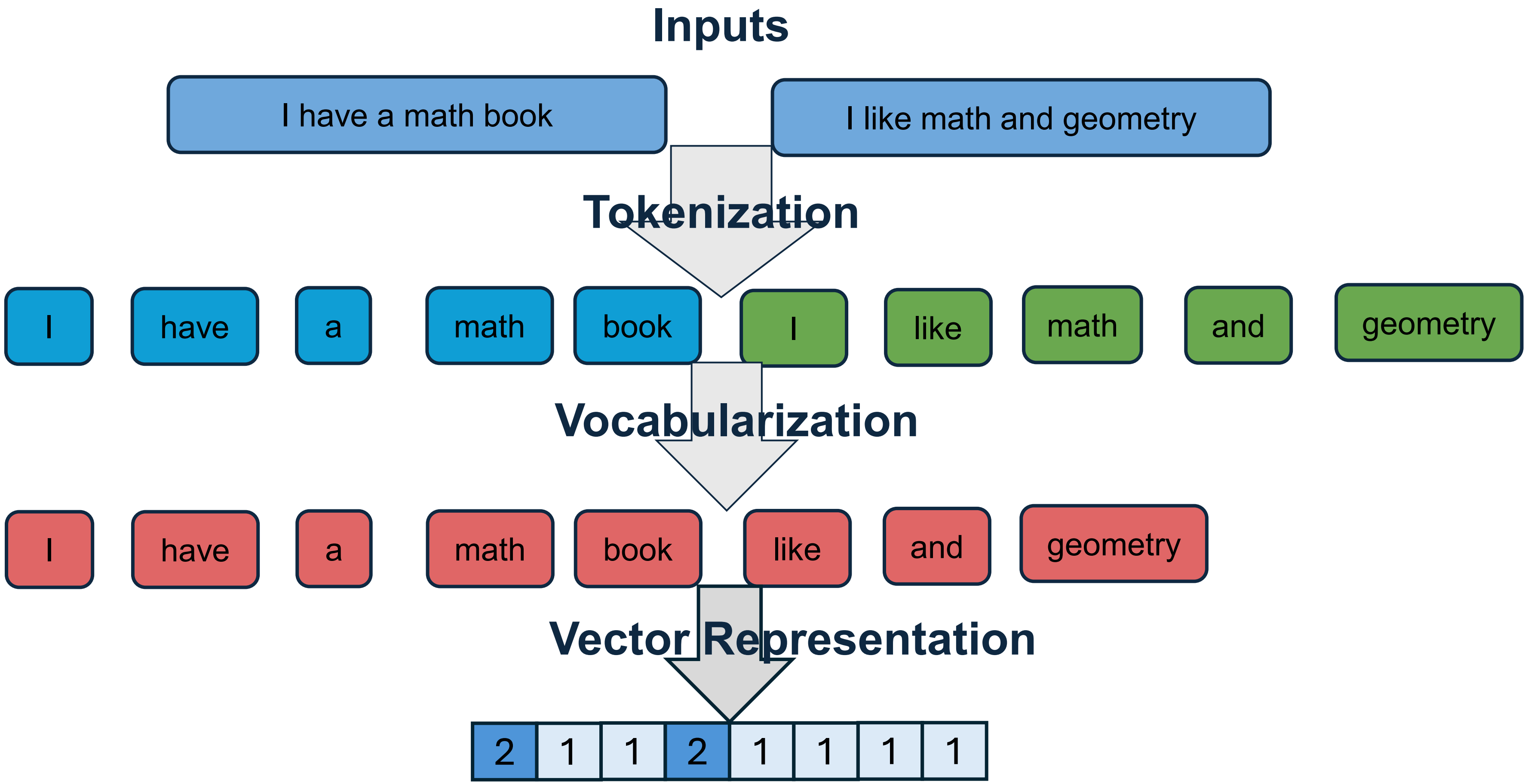
Qualitative data unlike numbers.

Lacking a specific format makes it difficult to analyze, compare, and manipulate.

Requires processing and interpretation.

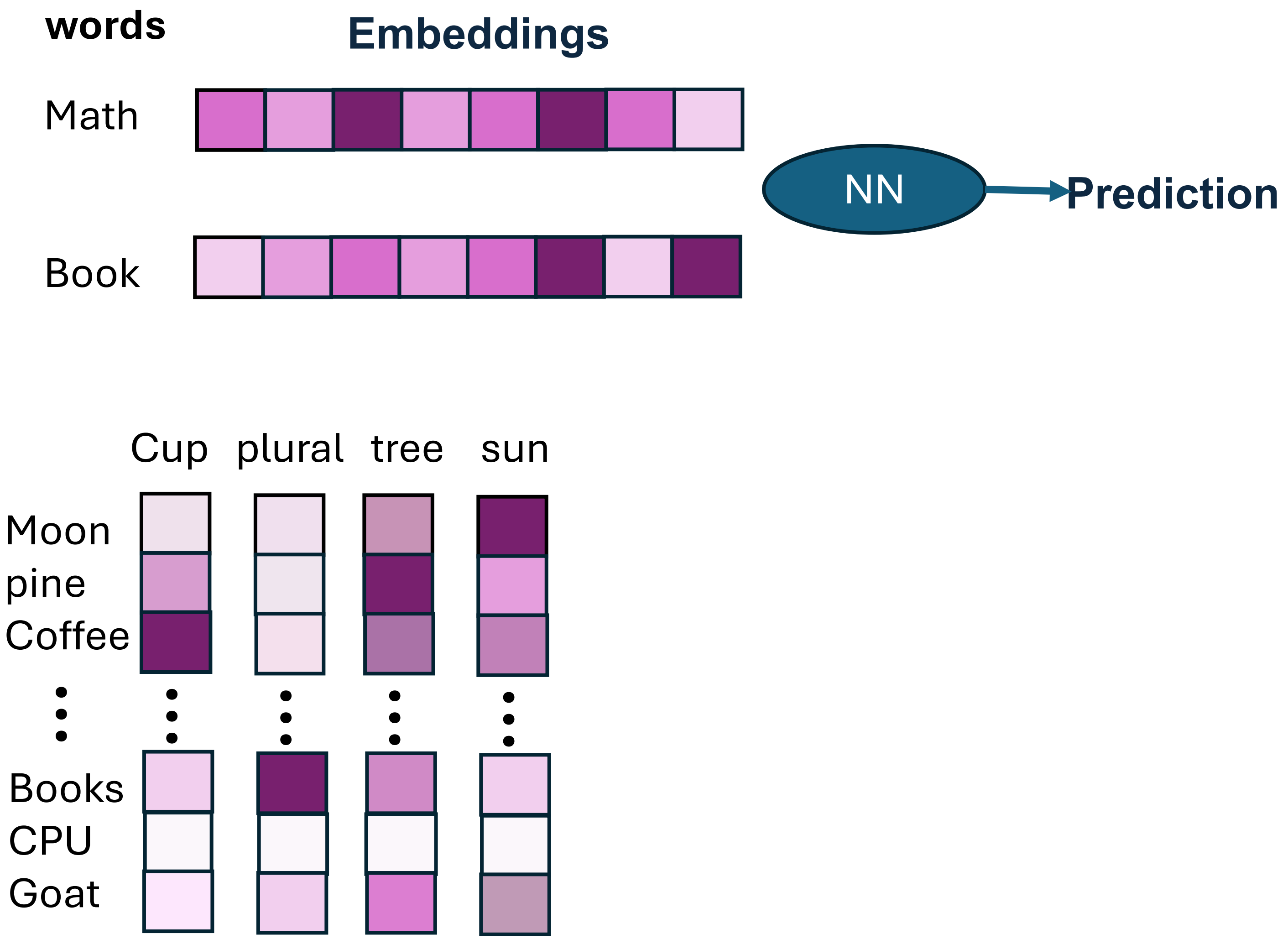
Bag of Words

- ❑ Tokenization: splitting up the sentences into individual words or subwords (tokens).
- ❑ combine all unique words to create a vocabulary used to represent the sentences.
- ❑ representations of text in the form of numbers.
- ❑ Each vector is a feature for the ML algorithm.



word2vec

- ❑ Word2vec: among early efforts to capture meaning of texts.
- ❑ Embeddings: vector representations of data.
- ❑ trained on huge amounts of textual data.
- ❑ neural networks generates word embeddings by comparing how often some words appear next to each other.
- ❑ How does word2vec capture meaning?



Encoding and Decoding

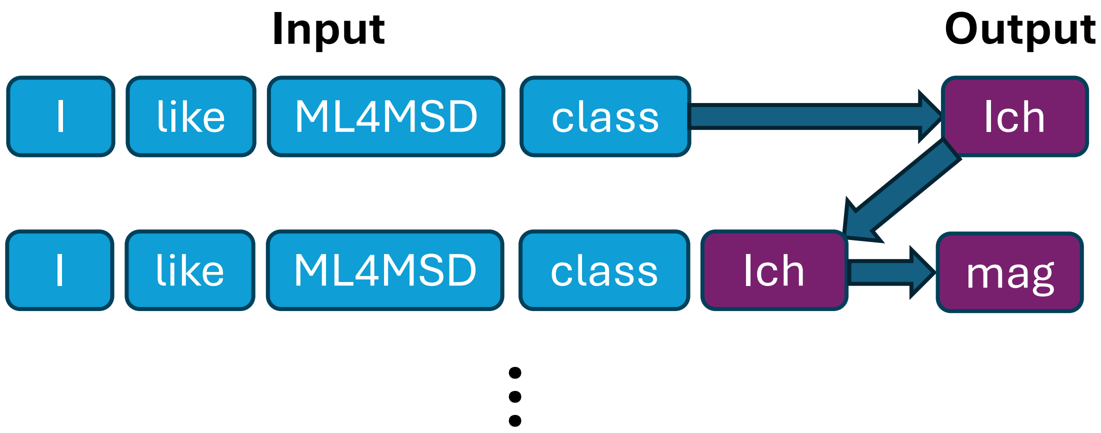
- Embeddings should vary by the context.

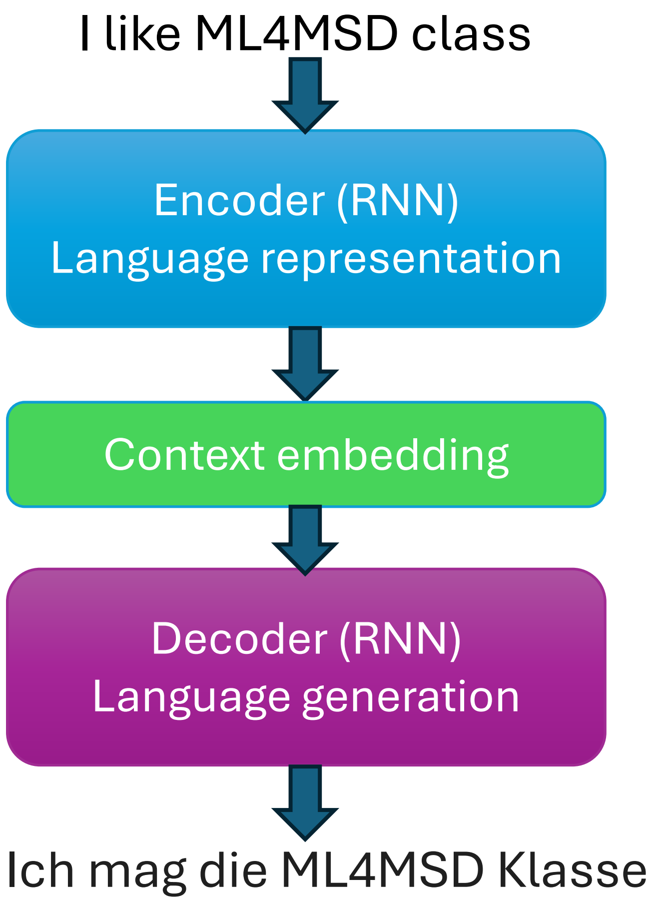
✓ Recurrent Neural Networks (RNN).

- RNNs are utilized for two tasks:

1. Encoding or representing the input
2. Decoding or generating an output sentence

❖ *Autoregressive steps.*





Attention Is All you Need

https://arxiv.org/pdf/1706.03762

Cited by 200714 as of 10/27/2025

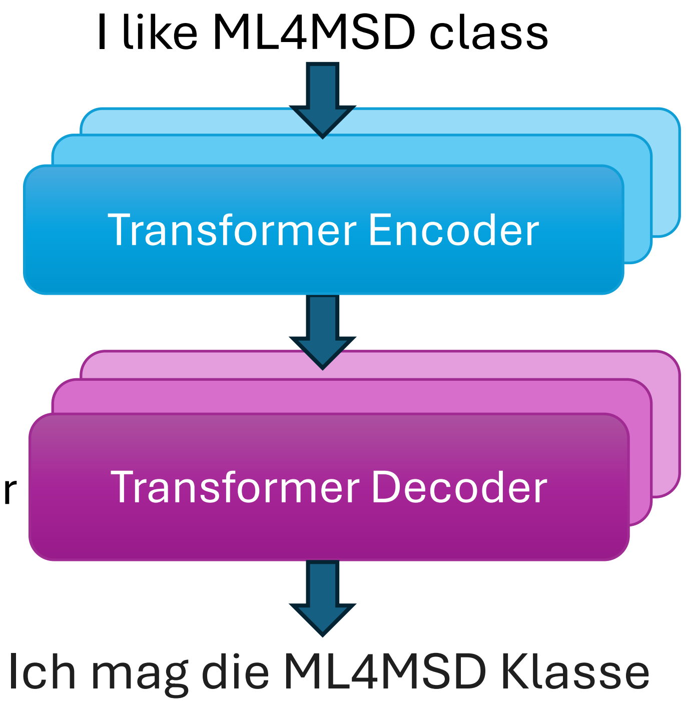
Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

This context embedding is a single embedding that represents the whole input.

It will be challenging to handle long sentences.

Solution? **Attention**



Attention focuses on parts of the input sequence that are relevant to each other

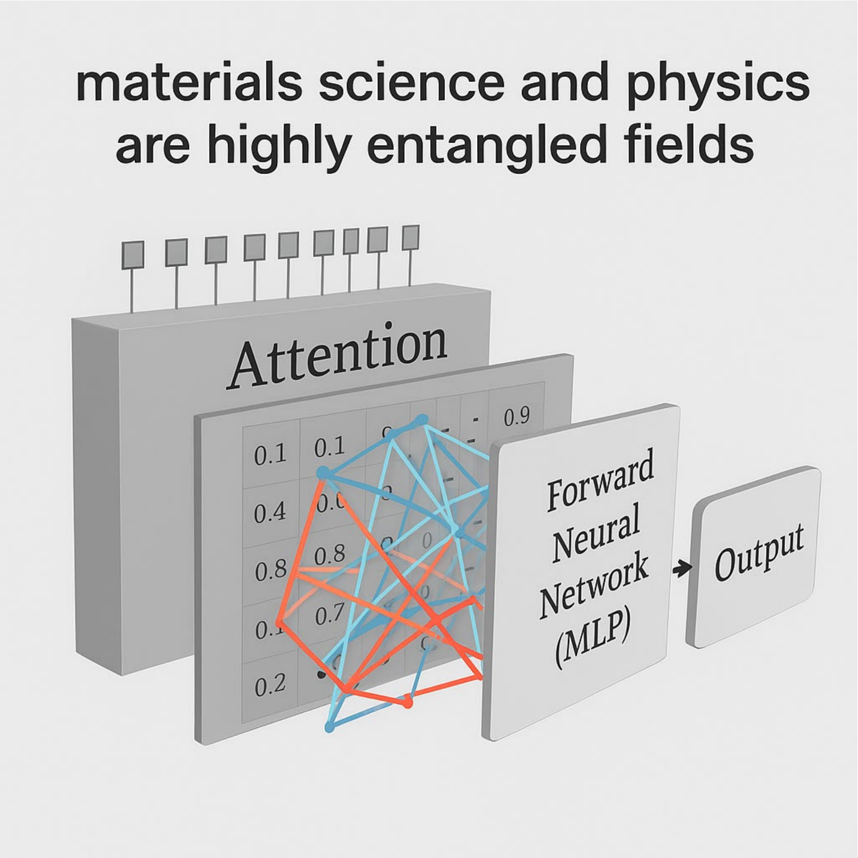
Attention determines which words are most important in a sentence.

Transformers (network architecture) based only on the attention mechanism

- ➢ **NO RNN**

Transformer

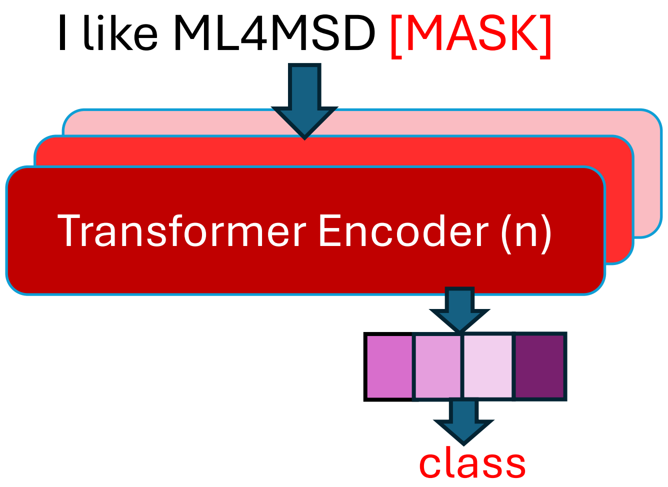
1. Tokenization.
2. Vector representation.
3. Attention mechanism.
4. Feed Forward Neural Network (FNN).
5. Probability distribution.
6. Repeating this process completes a text.

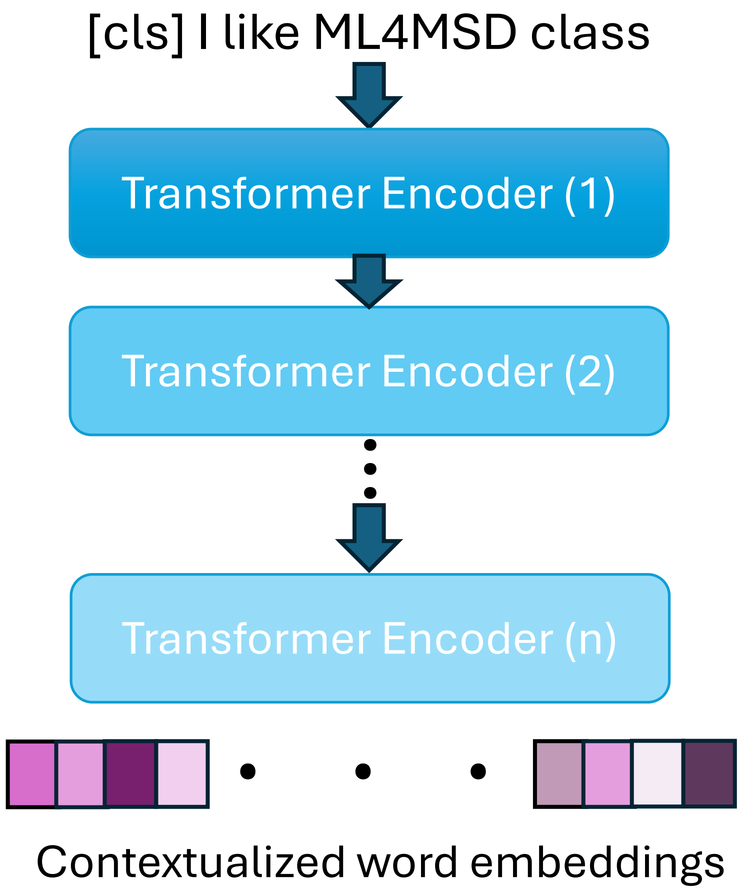


Encoder-Only Models: Representation Model

BERT : Bidirectional Encoder Representations from Transformers

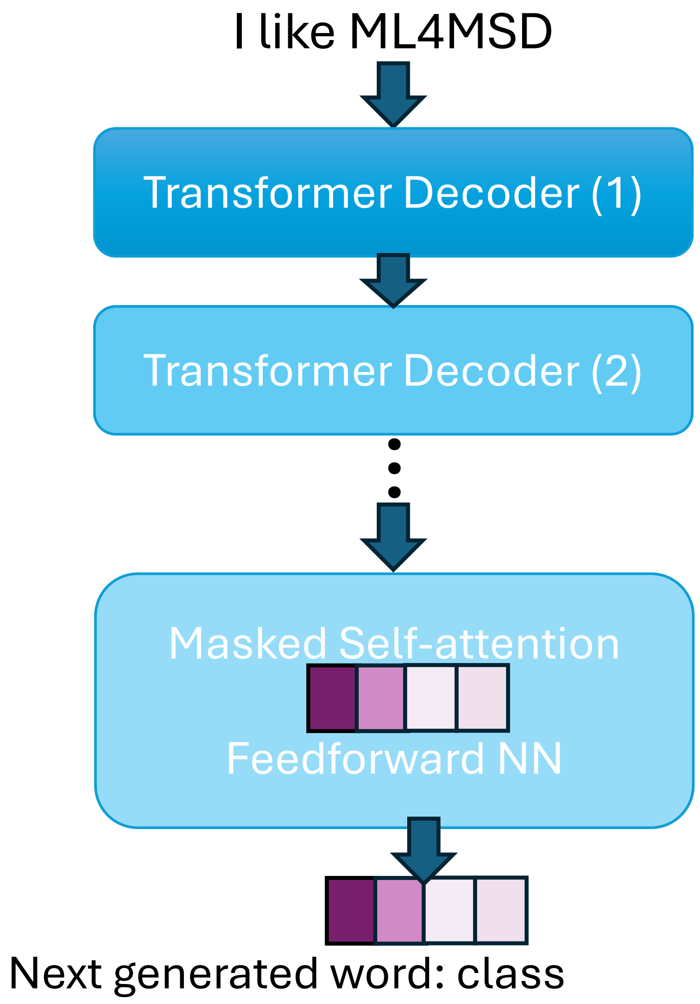
- Self-attention followed by feedforward neural networks.
- How to train encoder stacks ? *masked language modeling*
- Masking some words randomly in a text, then predicting the words based on the surrounding.





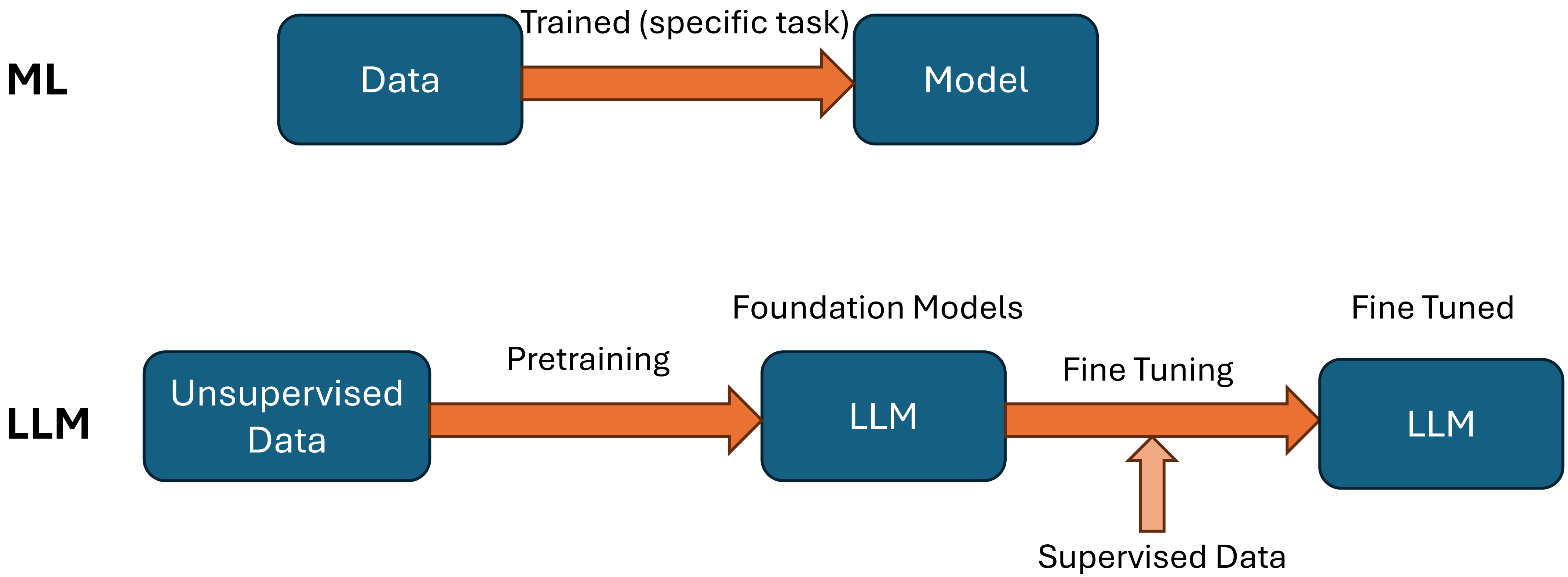
Decoder-Only Models: Generative Models

- **GPT** : Generative Pre-trained Transformer
- Generative LLMs: take in some text and attempt to autocomplete it.
- Can they be trained to answer us? **fine-tuning**
- Get a user query (*prompt*) and output a response.
- Generative models are *completion* models.



ChatGPT

LLM vs ML



ML4MSD_LLM_pt1 exercise

Applications of LLMs in Materials science

- ❖ Molecular and Material Property Prediction: LLM Spectrometry
- ❖ Molecular and Material Design: Data-Driven Design
- ❖ Automation and Novel Interfaces: microscope operations, DFT Parameters
- ❖ Scientific Communication and Education: Materials Science Teaching Assistant
- ❖ Research Data Management: Structured Data Directly from Speech
- ❖ Hypothesis Generation and Evaluation: Tree of Thoughts and Retrieval Augmented Generation
- ❖ Knowledge Extraction: Information extraction from literature

Knowledge extraction in Materials science

Knowledge extraction from literature.

Data is in the form of text, tables, figures etc.

Information extraction (IE) is the key factor in NLP to extract the relationships between named entities.

Traditional ML models need structured relationships between semantic entities of interest

e.g. $LiNbO_3$ is studied for ferroelectric

application

How is it implemented???

Information Extraction (IE)

What are the key entities in text for a Q&A task ? Information extraction (IE)

1. Extract entities: elements, crystal structure, formula... → **Named Entity Recognition (NER)**
2. Extract the relation between entities: application, description,... → **Relation Extraction (RE)**

Cu (name) electrodes were deposited using Physical vapor deposition technique. A 300-nm layer(description) of thermally grown silicon oxide (SiO_2) (formula) was used as the insulator(application) in our MOSFET (application).
Both token's location and label are needed!
Huge training resources like Wikipedia, books, websites,...

Entity recognition
Formula: Cu, SiO_2
Description: 300-nm layer
Application: electrodes, insulator, MOSFET
Name : silicon oxide

Coreference resolution
Entity A: silicon oxide (SiO_2)
Entity B: Cu
Entity C: 300-nm layer
Entity D: MOSEFET
Entity E: insulator
Entity F: electrodes

Relation Extraction
A,C: has_description
B,F: has_application
A,D: has_application
⋮

sequence-to-sequence

- Materials information may not always be modeled as simple pairwise relations!
- Depending on composition, morphology, crystal structure,…
- e.g., zinc oxide nanoparticles are catalysts, but “ZnO" and “nanoparticles" alone are not necessarily catalysts.

- Solution? **Encoder-decoder LLMs**

- LLMs are able to leverage semantic information between tokens in natural language sequences of varying length.
- A model is trained to output tuples of two/more named entities and the relation label belonging to the predefined set of possible relations between them.
- **Joint named entity recognition and relation extraction (NERRE)**

Document

Cu electrodes were deposited using Physical vapor deposition technique. A 300-nm layer of thermally grown silicon oxide (SiO_2)was used as the insulator in our MOSFET.

Output sequence

silicon oxide @NAME@ SiO_2@FORMULA@ @N2F@
300-nm layer @DES@ SiO_2@FORMULA@ @D2F@
Cu @FORMULA@ electrodes @APP@ @A2F@

LLM-NERRE

Hierarchical entity relationships without explicit enumeration.
LLM is fine-tuned to simultaneously extract named entities and their relationships.
Fine-tune a pretrained LLM to accept a text passage (for example, a research paper abstract) and write a precisely formatted “summary" of knowledge contained in the prompt.

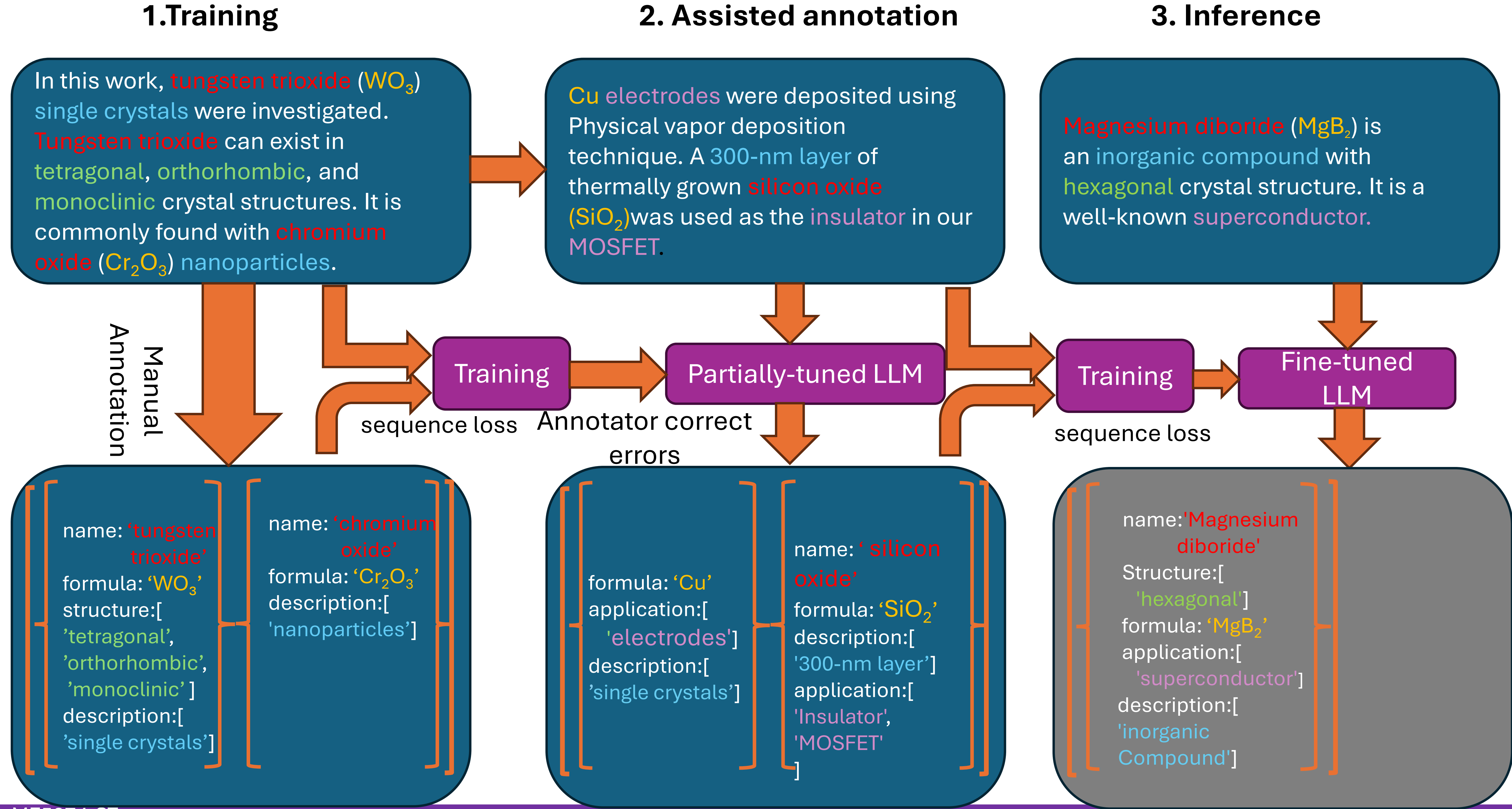
Document

Cu electrodes were deposited using Physical vapor deposition technique. A 300-nm layer of thermally grown silicon oxide (SiO_2)was used as the insulator in our MOSFET.

JSON documents

[Formula: 'Cu' Application: 'electrodes', 'MOSFET']

[Name: 'silicon oxide' Description: '300-nm layer' Formula: 'SiO_2' Application: 'electrodes', 'MOSFET']


1.Training
2. Assisted annotation
3. Inference
In this work, tungsten trioxide (WO_3) single crystals were investigated. Tungsten trioxide can exist in tetragonal, orthorhombic, and monoclinic crystal structures. It is commonly found with chromium oxide (Cr_2O_3) nanoparticles.
Cu electrodes were deposited using Physical vapor deposition technique. A 300-nm layer of thermally grown silicon oxide (SiO_2)was used as the insulator in our MOSFET.
Magnesium diboride (MgB_2) is an inorganic compound with hexagonal crystal structure. It is a well-known superconductor.
Manual Annotation
Training
sequence loss
Partially-tuned LLM
Annotator correct errors
Training
sequence loss
Fine-tuned LLM
name: 'tungsten trioxide'
formula: 'WO_3'
structure:[
'tetragonal',
'orthorhombic',
'monoclinic']
description:[
'single crystals']
name: 'chromium oxide'
formula: 'Cr_2O_3'
description:[
'nanoparticles']
formula: 'Cu'
application:[
'electrodes']
description:[
'single crystals']
name: ' silicon oxide'
formula: 'SiO_2'
description:[
'300-nm layer']
application:[
'Insulator',
'MOSFET'
]
name:'Magnesium diboride'
Structure:[
'hexagonal']
formula: 'MgB_2'
application:[
'superconductor']
description:[
'inorganic Compound']

MatBERT

A pretrained BERT model on materials science literature. MatBERT specializes in understanding materials science terminologies and paragraph-level scientific reasoning.

MatBERT_colab exercise

Patterns

CellPress
OPEN ACCESS

Article

Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science

Amalie Trewartha,[2,5] Nicholas Walker,[1,5,7,*] Haoyan Huo,[2,4,5] Sanghoon Lee,[1,4,5] Kevin Cruse,[2,4,5] John Dagdelen,[1,4,5] Alexander Dunn,[1,4,5] Kristin A. Persson,[3,4,6] Gerbrand Ceder,[2,4,6] and Anubhav Jain[1,6,*]

[1]Energy Technologies Area, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

[2]Materials Sciences Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

[3]Molecular Foundry, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

[4]Department of Materials Science and Engineering, University of California, Berkeley, 210 Hearst Memorial Mining Building, Berkeley, CA 94720, USA

[5]These authors contributed equally

[6]Senior author

[7]Lead contact

*Correspondence: walkernr@lbl.gov (N.W.), ajain@lbl.gov (A.J.)

https://doi.org/10.1016/j.patter.2022.100488

References

1. Hands-On Large Language Models: Language Understanding and Generation, Book by Jay Alammar and Maarten Grootendorst.
2. Attention is All You Need. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin.
3. Weston, L. et al. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature.
4. 34 Examples of LLM Applications in Materials Science and Chemistry: Towards Automation, Assistants, Agents, and Accelerated Scientific Discovery. Zimmermann et.al.
5. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. Trewartha, A. et al.
6. Structured information extraction from scientific text with large language models. Anubhav Jain et.al.
7. Efficient estimation of word representations in vector space. T Mikolov et.al.