Technology Innovation Institute

GICHD INNOVATION

The Evolutionary Journey of Large Language Models

From Early Foundations To Current Applications

Ahmed Alhammadi – Lead Researcher @ AICCU

Presentation Outline

- Evolution of Language Models
- Creating Falcon LLM
  - The Falcon Family of LLMs
- Fine-Tuning LLMs
- An Illustrative Example
  - Conventional Approach vs General LLM vs Fine-Tuned Legal LLM
- Challenges and Future Directions

The Evolution of Language Models - A Glimpse Into the Past

Foundation of Technologies:
• Understanding past advancements provides essential insights into current language model architectures and functionalities.

Appreciating Progress:
• Recognizing the timeline highlights remarkable growth and inspires continuous innovation in language processing.

Contextualizing Capabilities:
• Knowing the evolution clarifies the strengths and limitations of current models.

Inspiring Future Innovations:
• Historical insights motivate the development of more advanced, efficient, and versatile language models.

The Evolution of Language Models - A Glimpse Into the Past

Rule-Based Systems
(1950s)
Early NLP systems using handcrafted rules and dictionaries.

n-Gram Models
(1990s)
Statistical models predicting the next word based on previous n-1 words.

Feedforward Neural Networks
(2000s)
Neural networks with embeddings capturing semantic meanings.

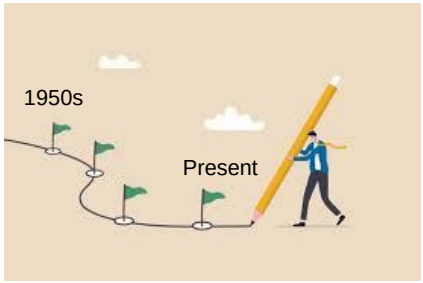
Recurrent Neural Networks (RNNs)
(1980s)
Networks with loops to handle sequential data and maintain hidden states.

Long Short-Term Memory (LSTM)
(1997)
RNN variant with gates to better capture long-term dependencies.

Gated Recurrent Units (GRUs)
(2014)
Simplified LSTM with fewer gates, easier to train.

Transformer Architecture
(2017)
Introduced self-attention mechanisms, enabling parallel processing.

Large Language Models (LLMs)
(2020s)
Massive models with billions of parameters trained on diverse datasets.



Classical Natural Language Processing Pipeline

1 Text Input

Raw text data to be processed.
{Sheikh Zayed was born on May 6, 1918, in Abu Dhabi}

2 Tokenization

Splitting text into individual words or tokens. Example:
{"Sheikh", "Zayed", "was", "born", "on", "May", "6", ",", "1918", ",", "in", "Abu", "Dhabi", "."}

3 Stemming /Lemmatization

Reducing words to their base or root form.
- Example:
- **"was" → "be"**
- **"born" → "bear"**

4 Part-of-Speech (POS) Tagging

- Assigning grammatical categories to each token. Example:
- **Zayed → NNP | Proper noun, singular**
- **born → VBN | Verb, past participle**
- **1918 → CD | Cardinal number**

5 Parsing (Syntax Analysis)

Analyzing the grammatical structure of sentences.

[S
[NP [NNP Sheikh] [NNP Zayed]]
[VP [VBD was] [VBN born]
[PP [IN on]
[NP [NNP May] [CD 6] [, ,] [CD 1918]]]
[PP [IN in]
[NP [NNP Abu] [NNP Dhabi]]]]

6 Named Entity Recognition

Identifying and classifying named entities in text.
Sheikh Zayed was born on **May 6, 1918**, in **Abu Dhabi**
- **Sheikh Zayed → Person**
- **May 6, 1918 → Date**
- **Abu Dhabi → Location**

7 Feature Extraction

Converting processed text into numerical features for modeling.
Example: **Bag-of-Words**, **Term Frequency - Inverse Document Frequency**

8 Modeling/Classification

The extracted features are used as input for machine learning models
Possible Applications:
- **Biography Generation**
- **Information Retrieval/Extraction**

Challenges Faced by Classical NLP

Challenge	Impact on Classical NLP
Ambiguity	Misinterpretation of meaning
Rule-Based Limitations	Poor scalability and adaptability
Extensive Feature Engineering	Time-consuming and requires domain expertise
Vocabulary Limitations	Failure to handle new or misspelled words
Multilingual Challenges	Difficulty processing multiple languages simultaneously

The Emergence of Large Language Models

What is an LLM?

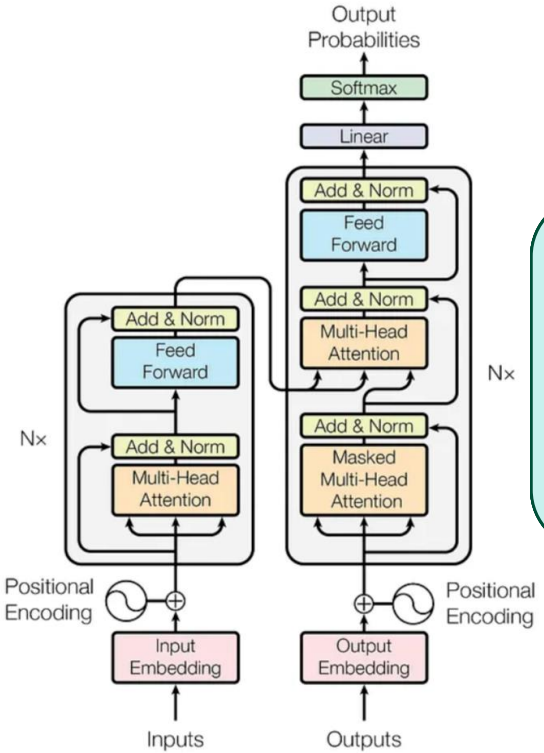
- Large Language Models (LLMs) are neural networks with billions of parameters.
- Trained on diverse, massive datasets to understand and generate human-like text.

More details:

- **Transformer Architecture:** Introduced in 2017 [1], it revolutionized NLP with self-attention, enabling efficient handling of long-range dependencies.
- **Self-Attention Mechanism:** Allows models to assess the importance of each word relative to others, capturing contextual relationships effectively.
- **Positional Encoding:** Integrates word order into embeddings, enabling Transformers to recognize sequence and contextual information within text.

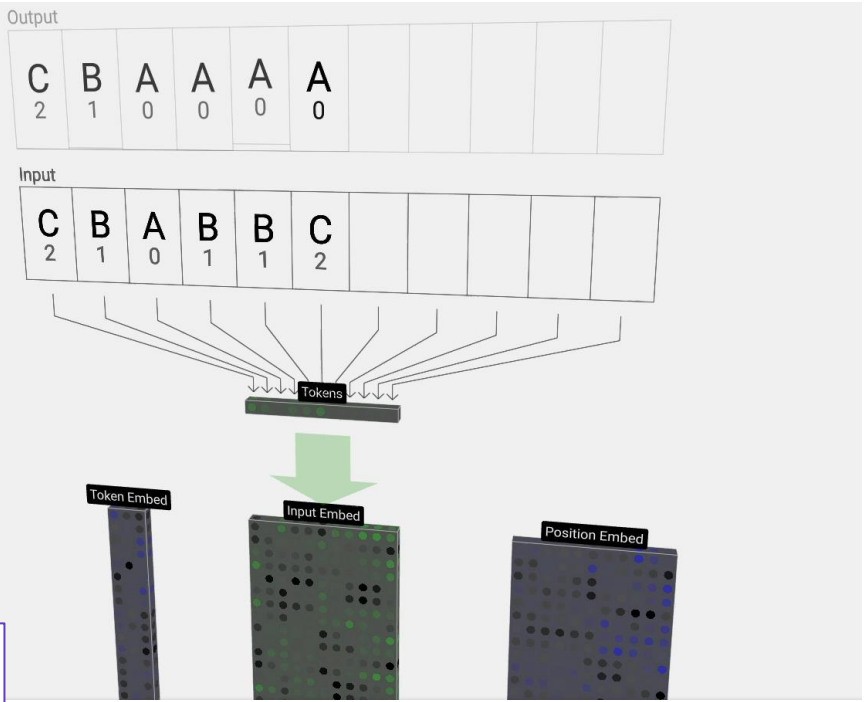
[1] M. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, pp. 6000–6010, Dec. 2017.

The Transformer – A Visual Explanation



Transformer Model Architecture
Source: M. Vaswani et al. "Attention is all you need," *Proceedings of the 31st NIPS*, Dec. 2017.

Generating Predictions
- At each position in the sequence, the model predicts the probability of each possible next token.
- 6th Entry Prediction:
- Model computes probabilities for 'A', 'B', or 'C' as the next token.
- Example Output: High probability assigned to 'A'.



Credit: https://bbycroft.net/llm

$$Attention(q, k, v) = \overbrace{softmax\left(\frac{qk^T}{\sqrt{d_k}}\right)v}^{\text{Attention weights}}$$

from (q) to (k, v); d_k: vector dimensionality of K, V

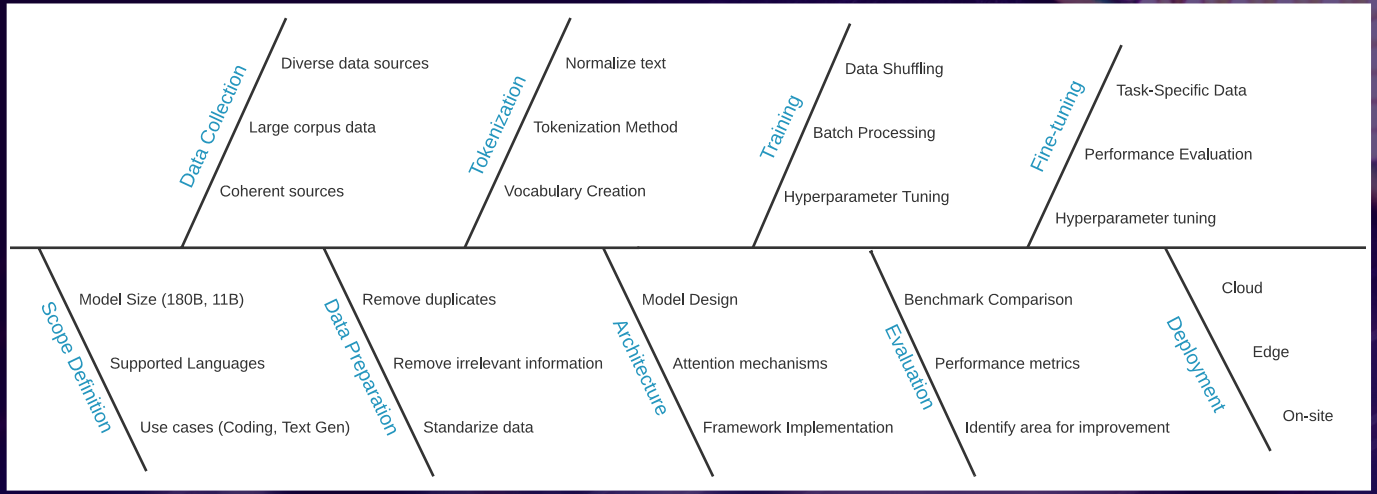
Classical NLP vs Large Language Models

Aspect	Classical NLP	LLMs
Feature Creation	Manual feature engineering	Automatic feature learning
Understanding	Limited context handling	Deep contextual understanding
Scalability	Challenges with handling large datasets	Designed to leverage massive datasets
Flexibility	Rule-based, less adaptable to new data	Highly adaptable and generalizable

Creating Falcon LLM

The process behind our models

The secret behind our success is the rigorous process we follow. Here is a high-level overview of the carefully crafted procedure we execute in order to achieve the best-performing models.



THE FALCON FAMILY OF LARGE LANGUAGE MODELS

- Generative LLMs offering unparalleled capabilities in natural language processing, understanding, and generation, tailored to a wide array of industries and applications.

Falcon-7B *Released in May 2023*	**Falcon-40B** *Released in May 2023*	**Falcon-180B** *Released in September 2023*	**Falcon 2-11B** *Released in May 2024*
7 billion parameters trained on 1.5 billion tokens	40 billion parameters trained on 1 trillion tokens	180 billion parameters trained on 3.5 trillion tokens	11 billion parameters trained on 5 trillion tokens

License: available under the permissive **Apache 2.0 software license**

Available on: Hugging Face, Vertex.ai, Amazon SageMaker, Azure Machine Learning

Falcon Family

- Summary

Model	Tokens	Availability	Supported Languages	Avg Performance Open LLM Leaderboard v1	Closest Model
Falcon-7B	1,500B	Apache 2.0	en, fr	44.17	<GPT-3
Falcon-40B	1,000B	Apache 2.0	en, cs, es, de, fr, it, nl, pl, pt, sv, ro	58.05	Chinchilla
Falcon-180B	3,500B	Responsible use license		67.85	PaLM-2 Large
Falcon2-11B LLM and VLM	5,500B	Responsible use license		64.28	Llama3-8B, Gemma

THE FALCON FAMILY OF LARGE LANGUAGE MODELS

- **Leading performance**

- Falcon2-11B VLM has surpassed, the top performing VLMs in their category
- Falcon2-11B has surpassed Llama-3-8B from Meta, becoming the top performing small-size pretrained LLM

	FALCON 2 11B	LAMA 38B (Pre-Trained)
ARC - C	59.73	**78.6**
HellaSwag	**82.91**	82.09
MMLU	58.4	**66.6**
TruthfulQA	**52.56**	43.9
WinoGrande	78.3	77.35
GSM8K	**58.37**	45.79
OVERALL	**64.28**	62.55

	Falcon2-11B VLM	LLaVA 1.6-Vicuna-13B	LLa VA 1.6-Mistral-7B
MME	**1589/343**	1575/326	1498/321
SQA	**74.9**	73.6	72.8
POPE	**88.4**	86.2	86.7
MV	37.2	35.1	**37.4**
VQAT	66.7	**67.1**	65.7
HB	**48.7**	44.5	41.7
Overall	**68.8**	67.7	65.3

- **Increasing adoption**

45+ million downloads of our LLM models

High media visibility

440M+ **impressions** on social media

11k+ **media articles** on outlets

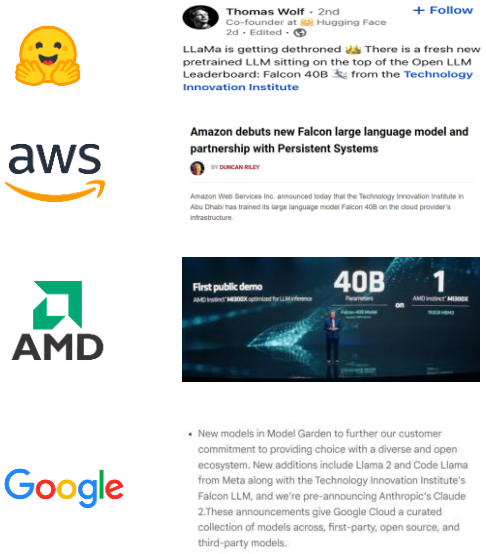
470k+ Falcon **website visits**

4M+ **engagements** on social media

(like, comment, share)

- **Recognition**

Officially recognized by multiple **large tech players**



Fine-Tuning Large Language Models

After building large language models through pre-training, what if we wanted the model **to excel in specific domains** such as legal, medical, or technical fields, how do we achieve this specialization?

An Effective Solution: Fine-Tuning

Fine-tuning a large language model offers:

- **Domain Expertise:** Integrates specialized knowledge unique to a field.
- **Enhanced Accuracy:** Improves precision and contextual relevance.
- **Cost-Effectiveness:** Reduces training time and resources compared to building a model from scratch.

Aspect	Pre-training	Fine-tuning
Definition	Training on a vast amount of unlabelled text data	Adapting a pre-trained model to specific tasks
Data Requirement	Extensive and diverse unlabelled text data	Smaller, task-specific labelled data
Objective	Build general linguistic knowledge	Specialize model for specific tasks
Computational Cost	High (large dataset, complex model)	Lower (smaller dataset, fine-tuning layers)
Training Duration	Weeks to months	Days to weeks
Example	Falcon 11B / GPT-4	FinGPT

Example Use Case: Fine-Tuning a LLM for Legal Applications

The Need for an LLM in the Legal Field

• **Complex Language**: Legal documents contain specialized terminology and intricate sentence structures.
• **High Volume of Documents**: Lawyers sift through vast amounts of text for case law, contracts, and regulations.

Aspect	Traditional Approach	LLM-Powered Approach
Time Consumption	Hours to days per contract	Seconds to Minutes per contract
Consistency	Varies by reviewer	High consistency across documents
Error Rate	Prone to human error	Reduced errors with continuous learning
Scalability	Limited by human resources	Highly scalable with computational resources
Cost	High labor costs	Lower operational costs after deployment

An Illustrative Example in The Legal Domain (1/3)

Scenario: Your company is entering into a data processing agreement with a new vendor. It's crucial to ensure that the contract complies with the General Data Protection Regulation (GDPR) to protect personal data and avoid legal penalties.

The Question:
"Does this data processing agreement comply with GDPR requirements for data protection and privacy?"

Conventional Approach (Manual Review by a Legal Professional)

Process:

- Review: A lawyer meticulously examines the contract for GDPR compliance.
- Time Required: Approximately **1 to 2 hours** per agreement, depending on complexity.
- Cost: High, considering legal fees which may range from **$300 to $800 per hour**.
- Outcome: Detailed, personalized legal advice tailored to your company's specific needs and risk profile.

An Illustrative Example in The Legal Domain. (2/3)

The Question:
“Does this data processing agreement comply with GDPR requirements for data protection and privacy?”

General LLM Response:
“The contract mentions data protection and privacy, which are important under GDPR.”

Limitations:

- **Superficial Analysis:** Acknowledges the presence of data protection clauses without assessing their adequacy.
- **Lacks Specificity:** Does not identify whether the clauses meet GDPR standards or highlight any deficiencies.
- **No Actionable Insights:** Fails to provide guidance on necessary improvements or compliance steps.

An Illustrative Example in The Legal Domain. (3/3)

The Question:
“Does this data processing agreement comply with GDPR requirements for data protection and privacy?”

Fine-Tuned LLM Response:
Upon reviewing the data processing agreement, the following GDPR compliance aspects are addressed:
• Data Processing Purpose: Clearly defined and limited to specified activities.
• Data Subject Rights: Provisions for data access, rectification, and erasure are included.
• Data Security Measures: Detailed technical and organizational measures are outlined to protect personal data.
• Sub-Processors: Requires vendor to obtain prior written consent before engaging any sub-processors.
Areas for Improvement:
• Data Breach Notification: The agreement should specify the timeline and procedures for notifying your company in the event of a data breach.

With Fine-Tuned LLM in Legal Domain, The response provided **Comprehensive Assessment** and **Actionable Recommendations**. It also **Highlighted Deficiencies**

Current Challenges in Large Language Models

- **Massive Computational Requirements**
  Advanced models demand enormous computational power, leading to high energy consumption and sustainability concerns.

- **Tendency to Hallucinations**
  LLMs continue to generate plausible but incorrect or nonsensical information, undermining trust and reliability.

- **Lack of Physical World Grounding**
  Models are ungrounded in physical reality, lacking an understanding of physics and real-world constraints.

- **Limited Common-Sense Knowledge:** LLMs often fail to apply everyday common sense, resulting in responses that overlook obvious real-world facts.

Future Directions for Large Language Models

- **Vision-Language Models Becoming Prevalent**
  Integration of visual and textual data enhances understanding and generation, enabling more comprehensive AI applications.

- **Rise of Vision-Language-Action Models**
  Combining vision, language, and action to create more interactive and dynamic AI systems for real-world applications.

- **Innovative Architectures for Efficient Training**
  New architectures like MAMBA offer linear complexity, significantly improving training efficiency and scalability.

- **Exciting Future Prospects**
  Rapid technological advancements make this an exhilarating time, with transformative impacts on various industries and daily life.


Technology
Innovation
Institute
Thank you!
Artificial Intelligence Cross-Center Unit (AICCU)
TII – Technology Innovation Institute
tii.ae