Questions:

1. **Learning the essential working of Large Language Models.**

What are LLMs?

Q1: Why do LLM's require large amounts of data for training? Would we achieve similar models if we use smaller amounts of data? What type of data is indeed crucial of all the data that is out there in the world?

- Large Language Models (LLMs) require vast amounts of data for training because they need to learn the intricacies of human language, including grammar, syntax, semantics, and context. Training on large datasets allows the model to capture a wide range of language patterns, making it more capable of understanding and generating text that resembles human communication.. Smaller datasets limit the model's capacity to generalize and may result in overfitting or lack of depth in understanding complex structures.
- No, the accuracy levels of the output provided by the model would be less and it can lead to potential biases in the model. High-quality, diverse, and representative data is crucial for training effective LLMs.

How do they get built?

Q2: Why is there a requirement for self attention mechanism, what advantage does it give? How does the presence of positional encoders affect LLMs. Mention a few points on what would happen if these two concepts were not implemented.

- The self-attention mechanism is essential because it allows LLMs to weigh the importance of different words in a sentence when generating or understanding language. This mechanism enables the model to capture long-range dependencies and relationships between words, which is crucial for tasks like translation or summarization where the meaning of a word depends on distant parts of a sentence.
- Positional encoders provide information about the order of words since the self-attention mechanism itself doesn't inherently consider word order.
- Without self-attention, the model would struggle to understand context beyond nearby words, leading to inaccurate predictions. Similarly, without positional encoders, the model wouldn't distinguish between sequences, treating sentences like a bag of words and losing the meaning conveyed by word order.

How do they work?

Q3: Describe in detail steps to build a transformer from scratch. Here be novel in terms of the architecture that you propose.

1. Data Preparation:
    - Collect and preprocess the dataset: To ensure the data is in a suitable format for training.
    - Tokenize: Break down the text into individual tokens (words or subwords).
    - Create input-output pairs: For tasks like machine translation, pair the input and output sequences.
2. Define the Model Architecture:
    - Encoder Stack:
        - Create multiple layers of self-attention and feed-forward neural networks.
        - Each layer applies self-attention to the input sequence, followed by a feed-forward network.
        - Residual connections and layer normalization are used for improved training stability.
    - Decoder Stack:
        - Similar to the encoder, but with masked self-attention to prevent the decoder from seeing future tokens during training.
        - The decoder also uses encoder-decoder attention to attend to the output of the encoder.
3. Implement Self-Attention:
    - Calculate query, key, and value vectors: For each token, compute these vectors using linear projections.
    - Compute attention scores: Calculate the dot product between the query vector of one token and the key vectors of all tokens.
    - Apply softmax: Normalize the attention scores using softmax to obtain attention weights.
    - Compute weighted sum: Multiply the value vectors by their corresponding attention weights and sum them.
4. Add Positional Encoding:
    - Create a matrix of positional encodings and add it to the input sequence.
    - The positional encodings can be learned or calculated using predefined formulas.

5. Training:
    - Define loss function: Use a suitable loss function, such as cross-entropy for sequence-to-sequence tasks.
    - Choose optimizer: Select an optimizer like Adam for efficient training.
    - Train the model: Iterate over the training data, adjusting the model's weights based on the calculated gradients.
6. Evaluation:
    - Evaluate on a validation set: Assess the model's performance using appropriate metrics (e.g., accuracy, BLEU score).
    - Fine-tune on a test set: If necessary, make minor adjustments to the model based on the validation results.

What is good about them?

Q4: List a minimum of 10 use cases of Generative AI that you see around and reason why they may be used. From a futuristic lens see what aspects of LLMs need to be retained and what to be developed.

1. **Text generation** – Enhances content creation for blogs, articles, and stories.
2. **Code completion** – Assists in faster software development.
3. **Chatbots** – Powers virtual assistants for businesses.
4. **Machine translation** – Translates languages instantly.
5. **Image generation** – Creates artistic designs or synthetic data.
6. **Speech synthesis** – Converts text to natural-sounding speech.
7. **Summarization** – Extracts key information from long documents.
8. **Sentiment analysis** – Analyzes emotions in text.
9. **Answering questions** – Retrieves information efficiently.
10. **Medical diagnosis** – Supports doctors by summarizing patient data and suggesting treatments.

- For the future, aspects like ethical AI and fine-tuning for bias removal should be retained and improved.

What is harmful about them?

Q5: Setup experiments to showcase the harm of LLM. Capture the outputs and describe what those are called technically.

- **Experiment** : Spreading Misinformation about political figures

- **Prompt** : Bots spreading misinformation on social media about a political figure hence affecting the outcome of an election.
- **Expected Output** : A number of bot account users interacting with genuine users by spreading false information about a certain political figure with the sole purpose of affecting the outcome of an election towards a candidate after his or her image was affected by false information.
- **Technical Term** : Fact-Checking Failure: LLMs can generate incorrect or misleading information, especially when trained on data that contains misinformation.

2. Learning the Current Inclusion of Languages

How many languages are included?

Q6: Write an essay about how languages develop and highlight why they might go extinct in some cases.

Languages are constantly evolving. They are born, grow, change, and sometimes die, shaped by a complex interplay of social, cultural, political, and historical forces. This dynamic process can be traced back to the earliest human communities. As groups migrated and interacted, their languages diverged, creating new dialects and eventually new languages. A good example is the migration of the Bantu people from South Africa to East Africa, which led to the emergence of various Bantu-speaking communities in the region.

Social and cultural factors significantly influence language development. As societies progress, their languages adapt, with new technologies, ideas, and experiences introducing new words and phrases. Language can also serve as a tool for reinforcing social identities and cultural values. For instance, slang used by Gen Z has led to the rise of resources like the Urban Dictionary, where new meanings and phrases are cataloged, often with meanings that deviate from their original definitions, sometimes in vulgar ways.

Political forces also shape language evolution. The rise of empires and nation-states often leads to the standardization of certain languages and the suppression of minority ones. This can impact language survival. A case in point is the dominance of Luganda in central Uganda, particularly in the capital Kampala, where Luganda is widely spoken even by non-Baganda people due to the influence of the Buganda Kingdom. Similarly, some languages dominate simply because of their larger number of speakers.

The extinction of languages is a serious cultural loss, representing the disappearance of heritage, knowledge, and diversity. To prevent this, efforts should focus on language revitalization, supporting multilingualism, and documenting endangered languages. Such efforts are crucial to ensuring that the rich tapestry of human languages continues to thrive for future generations.

Languages evolve through social interaction, trade, and cultural exchanges. However, they go extinct due to colonization, globalization, and the dominance of major languages over minority ones. Economic migration, political suppression, and lack of educational support further contribute to the disappearance of languages.

Why were they included?

Q7: Provide a timeline of events in history that provides evidence that certain languages were favored and some were not. What are the effects on trade and civilization?

- **1600s-1900s**: European colonization promotes English, French, and Spanish, marginalizing indigenous languages.
- **1945**: Post-WWII, English becomes the global lingua franca, further marginalizing non-colonial languages.
- **1980s-present**: Globalization, the rise of the internet, and international business favor English, Mandarin, and other major languages.

Who or which languages were excluded?

Q8: Write an essay on the impact of languages on society in terms of exclusion, what could be the reasons. If you had a chance to change history(just to save a language) what would you have done at that point of time?

The Impact of Languages on Society: Exclusion and Its Causes

Language is one of the most powerful tools for communication, shaping not only the way people interact but also how societies are structured. However, languages can also act as barriers, creating exclusion when certain groups are unable to participate fully in societal discourse due to linguistic differences. The exclusion caused by language manifests in various social, political, and economic dimensions, often leading to marginalization. In this essay, we will explore the impact of languages on societal exclusion, the reasons behind this exclusion, and what could be done to preserve and protect languages that face extinction.

Exclusion Through Language

Language can serve as a bridge, but it can also act as a gatekeeper. People who do not speak the dominant language of a society often find themselves excluded from opportunities in education, employment, and social integration. In multilingual societies, minority language speakers can be marginalized, unable to access services or participate in civic life because official communication takes place in a dominant language.

An example of this is seen in colonized societies, where the colonizers imposed their language as the official medium of communication, governance, and education. Indigenous languages were often relegated to informal use, leading to a slow erosion of linguistic diversity and the exclusion of native speakers from mainstream society. The legacy of colonization has left many former colonies with linguistic hierarchies, where speaking a colonial language such as English or French is equated with social mobility, while indigenous languages are marginalized, stigmatized, or even forgotten.

Reasons for Language-Based Exclusion

1. Colonialism and Language Imposition: Colonization played a significant role in imposing foreign languages on indigenous populations. In many parts of Africa, Asia, and the Americas, colonial powers replaced local languages with their own for administrative and educational purposes. This led to the suppression of native languages, creating a society where only those who spoke the colonial language could advance socially or politically.
2. Globalization: In today's globalized world, languages such as English, Mandarin, and Spanish dominate international business, media, and education. This has made it necessary for people to learn these global languages in order to access opportunities on a global scale. However, this has also led to the decline of smaller, regional languages, which are seen as less economically viable. As a result, speakers of minority languages can feel excluded from broader global interactions and left out of development opportunities.
3. Political Policies: National language policies can also contribute to exclusion. Governments that promote one language as the official language of a country may do so at the expense of others, especially in linguistically diverse regions. In countries like India, which has hundreds of languages, the promotion of Hindi as the national language has led to tension and feelings of exclusion among non-Hindi speakers.
4. Education Systems: The language of instruction in schools plays a critical role in either including or excluding students. When children are taught in a language

that is not their mother tongue, they can struggle to understand lessons, leading to lower educational outcomes. This linguistic barrier disproportionately affects children from minority language groups, deepening inequalities and reinforcing exclusion.

Changing History to Save a Language

If given the chance to change history to save a language, I would focus on implementing inclusive education policies during colonization. Instead of allowing colonizers to impose their languages on indigenous peoples, I would advocate for a bilingual or multilingual education system. This system would ensure that indigenous languages are taught alongside the colonizers' language, creating a balanced and inclusive approach to education. Had this approach been adopted, it would have safeguarded linguistic diversity while still allowing people to engage with the global languages imposed by colonial powers.

In this alternative history, indigenous languages would be formally integrated into government systems, media, and education, ensuring that they remain viable and respected in modern society. By elevating indigenous languages to the same status as colonial languages, societies would have developed a deeper appreciation for linguistic diversity, and the exclusion caused by language would have been mitigated.

Additionally, I would support cultural programs and literacy campaigns that prioritize the preservation and documentation of endangered languages. These programs would include resources for creating dictionaries, grammar guides, and literature in indigenous languages, ensuring that future generations have access to their linguistic heritage. By fostering pride in indigenous languages and promoting their use in public life, these languages would be preserved for centuries to come.

Conclusion

Language plays a crucial role in shaping societal interactions, but it can also lead to exclusion when linguistic diversity is not recognized or respected. The imposition of dominant languages through colonization, globalization, and political policies has marginalized speakers of minority languages, limiting their access to opportunities and contributing to their social exclusion. If history could be changed, implementing inclusive language policies during critical moments, such as the colonization era, could have preserved linguistic diversity and reduced language-based exclusion. By valuing and promoting all languages equally, societies can create more inclusive environments where linguistic diversity is celebrated rather than suppressed.

Why and what can be done?

Q9: Don the hat of language activist and provide action items for the scientific community to include languages which have historically been excluded. This language activist should give both generic action items which are applicable to any language in the world. Also, write an essay for your local community highlighting the importance of this crucial responsibility to save a language.

1. Promote linguistic diversity in schools.
2. Fund research in low-resource languages.
3. Provide digital tools for language learning.
4. Develop inclusive language policies in governments.
5. Translate vital documents into minority languages.

3. Learning the challenges of low resource languages

What are the challenges in data collection?

Q10: Write an essay on biases that you see in language Bari. If possible ask senior citizens how these biases crept in. List 10 steps that you would undertake to collect data that has least bias.

- 10 Steps:
    1. Involve native speakers.
    2. Collect data across dialects.
    3. Include all age groups.
    4. Avoid urban-centric language samples.
    5. Use neutral sources.
    6. Document language in natural settings.
    7. Record oral traditions.
    8. Cross-check data with community experts.
    9. Translate key phrases to prevent bias.
    10. Regularly update datasets.

What are the challenges in tokenization?

Q11: Compare and contrast subword tokenizer and character level tokenizer. Provide evidences of their applicability in different use cases

Subword tokenizers break down words into smaller units like prefixes or suffixes, handling rare or unknown words by splitting them into known subwords. In contrast, character-level tokenizers split text into individual characters, capturing fine-grained details but leading to longer sequences. Subword tokenizers are more efficient, with shorter sequences and better semantic representation, making them ideal for tasks like machine translation and pretrained models like BERT. Character-level tokenizers eliminate out-of-vocabulary issues and are useful in tasks with noisy inputs like speech recognition or text generation, but they come with higher computational costs due to longer token sequences.

Subword tokenization is used in machine translation systems like Google's Neural Machine Translation and models like BERT and GPT. Character-level tokenization has been applied in speech recognition and optical character recognition (OCR) tasks, where flexibility and handling misspellings or new words are crucial.

What are unique ways to overcome this?

Q12: Discuss about language families and how tokenization can benefit from it. Don the hat of a historian and check what are the languages similar. If you were to group languages that can benefit from having a same tokenizer, what would they be.

Language Families and Shared Features

Many languages within the same family have words with a common etymological origin. This means that tokenizers designed for languages within the same family could exploit these shared characteristics to improve tokenization efficiency and accuracy. For instance:

- Indo-European languages (e.g., English, German, French, Spanish, Russian) share significant grammatical structures and vocabulary. They are inflectional, meaning words change form to express tense, case, or number, but often retain recognizable subwords or stems.
- Sino-Tibetan languages (e.g., Mandarin, Cantonese, Tibetan) rely heavily on characters with unique meanings, which makes character-level tokenization particularly effective, as each character carries significant information.
- Niger-Congo languages (e.g., Swahili, Zulu, Yoruba) are agglutinative, meaning they form words by stringing together morphemes. Subword tokenization would work well here, capturing prefixes, stems, and suffixes effectively.

How Tokenization Can Benefit from Language Families

When tokenizing languages within the same family, certain rules and shared patterns can be leveraged to create more efficient models:

1. Morphological Similarities: Many languages in the same family share a similar word structure. For example, Romance languages like Spanish, Italian, and French all follow predictable patterns of conjugation and affixation. A subword tokenizer like Byte-Pair Encoding (BPE) can be used across these languages to break down words into meaningful subwords that recur across languages, reducing the size of the tokenizer's vocabulary.
2. Handling Inflection: In languages like German and Russian (both Indo-European), words undergo substantial inflection, often adding prefixes and suffixes. A tokenizer optimized for one language could be adapted for another, capturing these morphological changes without requiring a complete overhaul of the model.
3. Shared Syntax: Many languages in the same family also follow similar syntactical rules. Subword tokenizers benefit from this, as once a word is broken into stems and affixes, the overall sentence structure remains predictable.

Similar Languages and Tokenization Grouping

If I were to group languages based on shared linguistic characteristics that can benefit from having the same tokenizer, several clusters emerge:

1. Romance Languages (French, Spanish, Italian, Portuguese):
    - These languages share significant vocabulary and grammatical patterns due to their common Latin origin. A shared subword tokenizer would work well across all of them by capturing the frequent roots and affixes.
2. Germanic Languages (English, German, Dutch, Swedish):
    - These languages share similar syntax and word formation rules, making it possible to use the same subword or wordpiece tokenizer. Tokenizers can exploit recurring word formations.
3. Niger-Congo Languages (Swahili, Zulu, Yoruba):
    - Many of these languages use extensive affixation, with prefixes and suffixes attached to root words. Subword tokenization would capture these morphemes effectively, reducing the vocabulary size while accounting for the morphological complexity.
4. Turkic Languages (Turkish, Uzbek, Kazakh):
    - These languages are agglutinative, where words are formed by combining multiple morphemes. A shared subword tokenizer that handles affixes

efficiently would be useful in these languages, identifying stems and affixes across the Turkic family.

-

How have other low resource languages solved this issue?

Q13: Provide a detailed analysis of how other languages have done this. The expectation is to come up with a table, have columns about properties and rows with different languages.

Q14: This is the crux of your learning at this workshop.

Write a two page essay about this program, what you learnt and how it helped you think for Bari. Highlight your learnings about solving for Bari. Provide research to showcase this was not solved before and reason why it was not included. Use the lens of the future and write how this solution will be improved and what are the goto pointers to learn more.

4. Trying to solve for the missing parts

Data collection – solve a challenge by bringing or creating your data

Q15: Create a table about different types of data and their sources. Discuss the 3V - Volume Velocity Variety and how it affects the data that you have collected. Note what future data sources could be and how important safeguarding those sources are. Write a letter to the local mayor to come up with measures to unearth more data and to safeguard the data that is there publicly in terms of time.

Table: Types of Data and Their Sources

Type of Data	Source	Description
Structured Data	Databases, Spreadsheets	Organized in a predefined manner, easy to analyze (e.g., CSV files).

Unstructured Data	Social Media, Emails, Text Documents	No predefined structure, requires processing to analyze (e.g., posts, messages).
Semi-Structured Data	JSON, XML, HTML	Contains both structured and unstructured components, facilitating data interchange.
Sensor Data	IoT Devices, Wearables	Continuous data generated by sensors in real-time (e.g., temperature, humidity).
Transactional Data	E-commerce Platforms, Banking Systems	Records of transactions, typically structured and time-stamped (e.g., sales records).
Geospatial Data	GPS, Satellite Imagery	Data related to geographic locations, useful for mapping and analysis.
Audio/Video Data	Media Platforms, Surveillance Systems	Multimedia files requiring different processing techniques for analysis.

Discussion of the 3Vs

1. Volume:
    - Refers to the amount of data collected. In today's digital age, the sheer volume of data from various sources can be overwhelming, leading to challenges in storage, processing, and analysis. For instance, social media generates vast amounts of unstructured data daily, which requires robust data management strategies.
2. Velocity:

    - Describes the speed at which data is generated and processed. With real-time data sources like IoT devices and social media feeds, organizations must develop systems that can process and analyze data quickly to derive actionable insights. High-velocity data requires efficient processing algorithms to ensure timely decision-making.
3. Variety:
    - Encompasses the different types and formats of data available. The presence of structured, unstructured, and semi-structured data means that data processing systems must be adaptable to handle multiple data types effectively. Variety also poses challenges for data integration and consistency across platforms.

Future Data Sources and Their Importance

- Future Data Sources:
    - Enhanced IoT integration, increased digital engagement through augmented reality, and user-generated content through mobile applications.
    - Advances in data collection through wearable technology and smart cities will provide new streams of data that can be harnessed for various applications.
- Importance of Safeguarding Data Sources:
    - Safeguarding data sources is crucial to maintaining data integrity and ensuring the privacy and security of individuals. Implementing protective measures against unauthorized access and data breaches will encourage public trust and participation in data collection efforts.

Letter to the Local Mayor

Poni Henry,
Hai Malakal,
Juba, South Sudan.
ponihenry44@gmail.com
Friday, 27th September, 2024

Johnson Swaka
Juba, South Sudan.

Dear Mayor Hon. Johnson Swaka,

I hope this letter finds you well. I am writing to express my concern regarding the importance of data collection and safeguarding in our community. As we navigate an increasingly digital world, the ability to harness various types of data—from structured databases to real-time sensor data—has never been more critical for informed decision-making and resource allocation.

To effectively unearth more data, I propose that our city initiates community engagement programs that encourage residents to contribute local insights and information. Additionally, we should explore partnerships with local universities and tech organizations to establish data collection projects that tap into the wealth of information available in our community.

Moreover, safeguarding our existing data sources is essential. Implementing measures such as regular audits, cybersecurity training for personnel, and transparent policies regarding data access can ensure that our community's information remains secure and trustworthy.

By prioritizing these initiatives, we can foster a data-driven environment that not only enhances our city's operations but also promotes transparency and accountability. I would appreciate the opportunity to discuss this further and explore potential avenues for collaboration.

Thank you for considering this important matter.

Sincerely,
Poni Henry.

Data processing – solve the smallest component of this challenge

Q16: In 10 steps describe what techniques can be employed for data processing in this context. Compare and contrast with data processing of other languages. Do you think some technique has to be improved, if yes, how will you improve it? What does this improvement impact and also speculate the absence of this improvement and what it will lead to.

Tokenization & Vectorization – solve or at least suggest a solution at any level

Q17: What would happen if a vectorizer was absent, would you still be able to achieve the goal. Can you think of an innovative improvement component that can replace a tokenizer + vectorizer combination.

If a vectorizer were absent in a natural language processing (NLP) pipeline, achieving the goal of transforming text data into a format suitable for machine learning or deep learning models would be challenging. Here's a breakdown of the implications of not having a vectorizer and a potential innovative improvement component:

Implications of Absence of a Vectorizer

1. **Loss of Numerical Representation**: Without a vectorizer, textual data would remain in its raw form, making it impossible for algorithms that require numerical input to process the data effectively. Models such as logistic regression, support vector machines, or neural networks cannot operate directly on text.
2. **Inability to Capture Semantic Meaning**: A vectorizer plays a critical role in transforming text into feature vectors that capture the semantics and relationships between words. Without this, important contextual information and relationships would be lost, leading to poor model performance.
3. **Increased Difficulty in Feature Engineering**: Without a vectorizer, manual feature engineering would become necessary, which can be time-consuming and may not capture the complexities of language as effectively as automated methods.
4. **Reduced Model Effectiveness**: The absence of a vectorizer would limit the ability to leverage advanced modeling techniques, potentially leading to less accurate predictions or classifications. The overall effectiveness of NLP tasks, such as sentiment analysis, text classification, or entity recognition, would be compromised.

- Use existing structures and build a mock prototype of your solution (No code required to be learned or written)

Q18: create a high level design document and low level design document with block diagram architecture to capture the prototype.