**UNIVERSITY OF CAPE COAST**

**SCHOOL OF ECONOMICS**

****

**DEPARTMENT OF DATA SCIENCE AND ECONOMIC POLICY**

**COURSE: DATA CURATION AND MANAGEMENT**

**COURSE TITLE: DMA820**

**GROUP WORK**

**LECTURER: DR RAYMOND ELIKPLIM KOFINTI**

**GROUP MEMBERS**

|  |  |
| --- | --- |
| NAMES:  REGINA ROBERTSON  MOHAMMED KAMALIDIN  CLEMENT KWAKU BOADU  JOHN ABBIW BONNEY | INDEX NUMBERS  SE/DMD/24/0013  SE/DMD/24/0001  SE/DMD/24/0021  SE/DMD/24/0002 |

# ASSIGNMENT QUESTIONS

LMs: Language Models (LMs) have transformed the world of unstructured data. Present a brief history of LMs and demonstrate how their emerging abilities can revolutionise the field of data curation and management.

**TABLE OF CONTENT**

[ASSIGNMENT QUESTIONS 0](#_Toc205722488)

[LANGUAGE MODELS (LMS) 2](#_Toc205722489)

[INTRODUCTION 2](#_Toc205722490)

[ BRIEF HISTORY OF LANGUAGE MODELS 2](#_Toc205722491)

[ REVOLUTIONISING DATA CURATION AND MANAGEMENT THROUGH EMERGING LM ABILITIES 4](#_Toc205722492)

[ FUTURE POTENTIAL AND ETHICAL CONSIDERATION OF LMs 7](#_Toc205722493)

[ LIMITATIONS OF LANGUAGE MODELS 8](#_Toc205722494)

[ IMPACT OF LMs ON DATA CURATION AND MANAGEMENT 9](#_Toc205722495)

[CONCLUSION 11](#_Toc205722496)

[REFERENCES 12](#_Toc205722497)

# LANGUAGE MODELS (LMS)

## INTRODUCTION

Language Models (LMs) are advanced artificial intelligence systems designed to process, understand and generate human language in both written and spoken forms. They function by learning statistical and semantic patterns from vast amounts of textual data, enabling them to interpret context, infer meaning, and produce coherent responses or content. LMs are not limited to simple word prediction; they can perform complex linguistic tasks such as translation, summarization, question answering, sentiment analysis and dialogue generation. Their evolution reflects decades of progress in computational linguistics, natural language processing (NLP) and machine learning and their development can be traced through several key stages, from early rule-based systems to today’s large-scale, transformer-based architectures.

Many language models have developed over time, each marking significant advancements in the field. Vector-based word representations and context-aware embeddings were first introduced by pre-transformer era models like Word2Vec (2013), GloVe (2014) and ELMo (2018). With BERT (2018), the GPT series (2018–2025) and RoBERTa (2019), the transformer-based generation got underway and established new benchmarks for performance on a variety of NLP tasks. Capabilities were extended across languages and domains by multilingual and specialized models such as mBERT (2019), XLM-R (2020) and BioBERT (2019). PaLM 2 (2023), Claude 3 (2024), LLaMA 3 (2024) and Gemini 1.5 (2024) are some of the most recent and multimodal LMs that have advanced reasoning, safety, customization, and cross-modal understanding.

## BRIEF HISTORY OF LANGUAGE MODELS

The evolution of LMs reflects steady advances in computational linguistics, statistical modeling and deep learning and can be traced through four key stages. These stages are explained below.

1. **Rule-Based and Statistical Models (1950s-1990s)**

Early language models relied on rigid, handcrafted grammar rules, offering precision but lacking adaptability to varied inputs. In the late 1980s and 1990s, statistical methods such as *n*-gram models emerged, enabling systems to learn word probabilities from large text corpora. This shift, exemplified by IBM’s statistical machine translation models, marked a move toward probabilistic approaches that were more flexible and better suited to real-world language use, laying the groundwork for modern language modelling. An n-gram model predicts the next word based on the previous *n–1* words. For example, a bigram model (n=2) trained on “The cat sat on the mat” extracts word pairs like *“The cat”* and *“cat sat”*, counts their frequencies, and calculates probabilities (e.g., P(cat | The) = 1.0). Larger datasets and higher-order n-grams capture richer context, with smoothing used to handle unseen combinations. This probabilistic approach replaced rigid grammar rules, learning from real text usage.

1. **Neural Network Language Models (2000-2014)**

The early 2000s marked the beginning of a new era in language modelling with the introduction of neural network-based approaches. Neural network language models replaced *n*-gram limitations with continuous vector representations, treating words as points in high-dimensional space to capture semantic relationships. Breakthroughs like Word2Vec (2013) and GloVe (2014) enabled context-based word embeddings, improving tasks such as classification, sentiment analysis, and semantic search. This shift provided greater flexibility and generalization, paving the way for the deep learning advances that followed. Example, each word is represented as a vector in high-dimensional space. Words like “good,” “wonderful,” and “fantastic” end up close together because they appear in similar contexts.  
When the model sees “absolutely wonderful,” it recognizes “wonderful” is near “good” in vector space, so it correctly predicts the review sentiment as positive even without having seen the phrase before.

1. **Deep Learning and Sequence Models (2015–2017)**

The period from 2015 to 2017 saw rapid advances in deep learning architectures that significantly improved the handling of sequential data. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, improved the ability to capture long-term dependencies in text, supporting tasks like translation and summarization. Sequence-to-sequence architectures enhanced neural machine translation, while attention mechanisms improved accuracy and interpretability by focusing on relevant input elements. These advances bridged early neural models and the transformer architecture, forming the foundation for modern Natural Language Processing (NLP) breakthroughs. Example, in 2017, advances like LSTMs, sequence-to-sequence models and attention mechanisms enabled AI meeting assistants to produce concise, accurate summaries of hour-long meetings, preserving key details, focusing on important points and outperforming earlier RNN-based approaches.

1. **The Transformer Era (2018 - Present)**

A decisive turning point arrived in 2017 with the publication of “Attention Is All You Need” by Vaswani et al. (2017), introducing the transformer architecture. Transformers revolutionized NLP by replacing sequential processing with self-attention, enabling faster training on massive datasets. BERT (2018) set new performance benchmarks through bidirectional context learning, while OpenAI’s GPT series advanced generative capabilities, with GPT-2 through GPT-5 achieving high fluency, reasoning skills, and multimodal processing. Optimized variants like RoBERTa, multilingual models such as mBERT and XLM-R, and domain-specific models like BioBERT broadened applications, while recent frontier Large Language Models (LLMs) like PaLM 2, Claude 3, LLaMA 3, and Gemini 1.5 introduced enhanced reasoning, safety, and multimodal integration. Today, transformers and LLMs are central to information retrieval, content generation and AI assistants, making them critical for automating metadata creation, analyzing unstructured data, and delivering real-time insights in data curation and management. Example, transformers like BERT, GPT-4 and LegalBERT enable law firms to analyze lengthy legal documents in seconds, generating summaries, extracting key clauses, and identifying risks cutting review time, improving accuracy and freeing staff for higher-value tasks.

## REVOLUTIONISING DATA CURATION AND MANAGEMENT THROUGH EMERGING LM ABILITIES

Recent advances in Language Models (LMs) have introduced capabilities that extend far beyond basic language understanding, making them transformative tools for the field of data curation and management. By leveraging their ability to process, interpret and generate human-like text at scale, LMs address long-standing challenges in handling large, complex, and often unstructured datasets. Below is an outline of how the emerging abilities of Language Models (LMs) are revolutionising data curation and management:

1. **Data Governance and Compliance**

Emerging LMs are transforming how organizations handle data governance by providing automated, intelligent solutions for maintaining data quality, security and regulatory compliance. One of their most impactful capabilities lies in the automatic detection and redaction of sensitive information from large, unstructured datasets, ensuring adherence to data protection regulations such as the General Data Protection Regulation (GDPR) and other jurisdiction-specific laws. LMs can be integrated into data pipelines to scan documents, emails, chat logs and databases in real time, flagging or removing personally identifiable information such as names, addresses, phone numbers, social security numbers, and biometric identifiers. This automation drastically reduces manual review workloads, minimizes human error and accelerates compliance audits. Example, in the financial sector, a bank leverages an LM-powered system to process thousands of customer emails daily. The model automatically identifies and redacts sensitive data such as social security numbers, account details and credit card information before the messages are archived or shared with other departments. This not only safeguards customer privacy but also ensures the bank remains fully compliant with GDPR requirements and internal data governance policies.

1. **Automated Metadata Generation**

Creating metadata has traditionally been slow and labour-intensive, relying on human cataloguers to manually interpret and label resources. Language Models transform this process by rapidly analysing large volumes of unstructured content and generating accurate, context-aware tags. For instance, a model like RoBERTa tags research papers with keywords (e.g., “machine learning”, “neural networks”) for a university’s digital library, improving discoverability. Also, the Ghana Statistical Service could use an LM to instantly tag and classify newly collected survey responses such as household demographic data or agricultural census forms as they are uploaded from field enumerators’ devices. This would enable immediate organisation, faster quality checks and quick access for analysts, greatly accelerating the time from data collection to actionable insights. This automation cuts costs, speeds up curation and greatly improves the discoverability of collections, making them more accessible to researchers and the public.

1. **Semantic Search and Discovery**

Semantic search and discovery capabilities allow users to retrieve information based on meaning rather than exact keyword matches. This is particularly valuable for curators managing multilingual or domain-specific collections, as LMs can bridge vocabulary gaps, identify conceptual equivalents, and surface relevant materials that traditional search engines might miss. For example, in a global climate change research repository, an LM-powered search could return relevant studies when a user searches for “sea level rise,” even if the original documents use related terms such as “coastal flooding,” “ocean inundation,” or the French phrase *élévation du niveau de la mer.* Again, A legal firm uses a Transformer-based model to retrieve case law documents relevant to “breach of contract” queries, even if exact phrases are absent.

1. **Data Cleaning & Standardization**

LMs contribute to data cleaning and standardization by detecting anomalies, filling in missing fields and harmonizing inconsistent metadata formats. This ensures that datasets meet interoperability standards required for integration into larger, cross-institutional repositories critical for collaborative research and policy-making. For example, in a multi-university research project on public health, an LM could automatically identify and correct inconsistencies in hospital data where some records use “M” and “Male” interchangeably for gender, fill in missing country codes based on location names and standardize date formats across all datasets to meet agreed-upon metadata standards. Moreover, GPT-4 standardizes inconsistent date formats (e.g., “12/01/23” vs. “January 12, 2023”) in medical records and extracts patient data into JSON for database integration.

1. **Automated Summarization**

Automated summarisation tools powered by LMs can distil lengthy datasets, reports or meeting minutes into concise, actionable summaries. This accelerates decision-making processes, particularly in high-volume data environments where manual review is impractical. GPT-3 summarizes 50-page quarterly reports for a financial analyst, highlighting key metrics like revenue growth and operational risks. For example, at the Ghana Statistical Service (GSS), an LM could automatically summarise extensive census reports or quarterly economic surveys into short briefs highlighting key demographic trends, shifts in employment patterns or changes in inflation rates. These summaries could then be quickly shared with policymakers, ministries and the media, ensuring timely and informed decision-making without the delays caused by manually sifting through large volumes of statistical data.

1. **Natural Language Querying**

Natural language querying allows non-technical users to interact with databases through conversational questions instead of structured query languages like SQL. This democratizes access to curated data, enabling educators, policymakers, and community stakeholders to extract insights without specialised technical training. A retail company’s chatbot, powered by a conversational LM, answers “Show me all late shipments in Q2 2025”, querying a logistics database in real-time. For example, at Akosombo VRA JHS No. 2, a teacher could simply type or speak a question such as, *“Show me the average math score for Form 2 students last term”* or *“List all students with attendance below 80% this year”*. The LM-powered system would instantly retrieve and present the results in a clear, easy-to-read format. This eliminates the need for staff to understand complex database commands, speeds up information access, and empowers the school to make timely, data-driven decisions about academic support and resource allocation.

1. **Predictive & Generative Analytics**

Predictive and generative analytics extend the role of LMs into proactive data management. By identifying emerging trends, forecasting data patterns or even generating synthetic datasets for testing and modelling, LMs support forward-looking strategies that anticipate information needs rather than merely responding to them. A healthcare provider uses GPT-4 to generate synthetic patient histories for AI model training, avoiding privacy violations under HIPAA. For example, the Ghana Statistical Service could use an LM to forecast population growth trends across regions based on historical census and migration data, helping guide infrastructure and resource planning. At Akosombo VRA JHS No. 2, an LM could analyse several years of student performance data to predict which students are at risk of underperforming in the final exams, enabling teachers to implement targeted interventions early. Additionally, LMs could generate realistic synthetic student datasets for training new school data management staff without exposing actual student information, ensuring both effective capacity building and data privacy.

## FUTURE POTENTIAL AND ETHICAL CONSIDERATION OF LMs

As Language Models (LMs) continue to evolve, their integration with complementary technologies such as knowledge graphs, semantic web tools and real-time data sources (e.g., DeepSearch or streaming APIs) is expected to further revolutionize data curation. By combining the contextual reasoning capabilities of LMs with the structured relationships in knowledge graphs, systems can provide richer, more accurate and interconnected insights. Similarly, the ability to pull and process up-to-the-minute information from the live web can significantly enhance the timeliness and relevance of curated datasets, supporting dynamic decision-making in rapidly changing domains such as finance, healthcare and climate monitoring.

However, alongside these opportunities lie important ethical and operational challenges. One of the most prominent is bias, which can manifest in subtle ways, such as reinforcing gender, racial or cultural stereotypes present in the training data. If left unchecked, such biases can distort analytical outputs and perpetuate systemic inequalities.

Another concern is interpretability as LMs grow in complexity, it becomes increasingly difficult to explain their reasoning or verify their accuracy, raising questions about accountability in high-stakes applications. Privacy also remains a critical issue, especially when models are exposed to or capable of inferring sensitive personal data. Without proper safeguards, the misuse or unintended leakage of such information could have serious consequences.

Addressing these risks requires transparent model design that documents training data sources, model architecture and known limitations. Moreover, implementing robust data governance frameworks including clear policies on data access, consent, anonymization and audit trails is essential to maintain trust and compliance with legal standards such as the General Data Protection Regulation (GDPR). As LMs become more embedded in decision-making pipelines, multidisciplinary oversight involving technologists, ethicists, policymakers, and domain experts will be critical to ensuring that these technologies are deployed in ways that are both innovative and socially responsible.

## LIMITATIONS OF LANGUAGE MODELS

Until recently, language models were riding a wave of intense hype but that buzz now seems to be fading. Like most new technologies, they’re settling into a more realistic, less sensational phase. For example, Klarna, the buy-now-pay-later platform, cut over 1,000 international staff as part of a significant shift towards AI, spurred by its 2023 partnership with OpenAI. However, in Spring 2025, the Swedish company admitted that its heavy reliance on AI-powered chatbots for customer service, which nearly halved its workforce in two years, led to quality issues and lower customer satisfaction. Recognising that most customers prefer human interaction, Klarna has now started re-hiring. Below are some of the limitations of LMs.

1. **Hallucinations**

A critical limitation of language models is hallucination, where they generate convincing but false information. Even advanced systems like GPT-4 average only 78% factual accuracy, far below the reliability needed for high-stakes domains such as healthcare, law and journalism. Real-world incidents such as fabricated Supreme Court cases, false statements about public figures and misinformation in app summaries highlight the risks when curated datasets or AI-generated outputs are not rigorously verified. The root causes lie in LMs’ inability to distinguish truth from falsehood, lack of self-fact-checking and the “helpfulness trap,” where the model prioritizes user satisfaction over accuracy. These hallucinations are not just programming errors but inherent to the models’ statistical design, making perfect factual accuracy theoretically unattainable. For data curation, this underscores the non-negotiable need for human oversight, cross-referencing with trusted sources and robust validation pipelines before integrating AI outputs into critical systems. Examples include; an AI summary tool misinformed some users of the BBC Sport app that tennis star Rafael Nadal had come out as gay and that he’s Brazilian (neither of these is true). Also, the error originated from Apple Intelligence, Apple's new AI software, launched in the UK in December 2024. One of its features is providing users with a concise summary of their missed app notifications. Well, sometimes (rather often) it made overly bold assumptions, so Apple had to suspend the alerts on iPhones.

1. **Language Models Fail in General Reasoning**

No matter how advanced the AI model is, its reasoning abilities lag behind big time. This includes common-sense reasoning, logical reasoning and ethical reasoning. Apple’s recent study, The Illusion of Thinking, explores how advanced models that generate detailed reasoning steps (like Claude 3.7 Thinking or DeepSeek-R1) fail surprisingly. In controlled puzzle environments, these so-called “reasoning models” improve on medium tasks compared to less advanced models, but collapse entirely as complexity increases.

Even more perplexing, as problems get harder, models start thinking less, cutting off reasoning before reaching any valid solution. This happens despite plenty of available compute, revealing a hard limitation not in hardware but in reasoning design. “These models fail to develop generalizable problem-solving capabilities for planning tasks, with performance collapsing to zero beyond a certain complexity threshold,” the authors write. The models also often “overthink” simple problems, continuing after finding a correct answer, which wastes resources and can introduce errors.

1. **Lack Transparency in Explaining Decisions**

Language models can’t explain their way of thinking. The "Chain-of-Thought" (CoT) prompting technique, where LMs display their step-by-step reasoning, has been considered a valuable tool for AI safety and interpretability. However, a critical Anthropic study reveals that CoT may not accurately reflect the LM's “thought process”.

The researchers gave the models a question along with a hint beforehand, then asked them to explain how they reached their answers. Claude 3.7 Sonnet referred to the hint only 25 percent of the time, while DeepSeek R1 mentioned it 39 percent of the time. Like a student caught cheating, models can bluff their way through an explanation.

1. **May Generate Biased or Offensive Content**

Due to the presence of biases in training data, LMs can negatively impact individuals and groups by reinforcing existing stereotypes and creating derogatory representations, among other harmful consequences.

Even models aligned with human values and appearing unbiased on standard benchmarks can still harbor widespread implicit stereotype biases, like a university professor, a staunch advocate for diversity and inclusion, consistently assigning all "organizational tasks" (like scheduling meetings and taking notes) to his female graduate students, although there are roughly as many men as women in the faculty.

## IMPACT OF LMs ON DATA CURATION AND MANAGEMENT

Table 1: Impact of LMs on Data Curation and Management

| **Challenge in Data Curation** | **LM-Driven Solution** | **Impact** |
| --- | --- | --- |
| Large volumes of unstructured data | Automated classification & tagging | Faster organization & retrieval |
| Inconsistent metadata standards | Context-aware metadata harmonization | Improved interoperability |
| Manual, slow data cleaning | Intelligent anomaly detection | Higher accuracy & efficiency |
| Limited access for non-technical staff | Natural language database interfaces | Broader participation in data usage |
| Difficulty spotting trends | Predictive analytics & pattern detection | More informed decision-making |

Source: Authors’ construct based on data from online.

From the table 1 above, Language Models (LMs) have brought significant transformations to data curation and management by addressing long-standing challenges in the field. One of the key issues in data curation is the overwhelming volume of unstructured data, including text, images and multimedia. LMs help overcome this challenge through automated classification and tagging, enabling vast amounts of content to be systematically categorized and labeled for quick retrieval. This leads to faster organization of datasets and minimizes the time spent on manual cataloguing.

Another persistent challenge is the inconsistency in metadata standards, which can hinder smooth integration between systems. By leveraging context-aware metadata harmonization, LMs can align and standardize metadata fields based on semantic understanding, ensuring that datasets from different sources follow a unified structure. This significantly improves interoperability and facilitates seamless data sharing across platforms.

The process of cleaning data has traditionally been manual and time-consuming, often prone to human error. LMs address this problem through intelligent anomaly detection, identifying errors, duplicates and missing values by recognizing patterns that deviate from the norm. In many cases, they can also suggest or perform automated corrections, resulting in higher accuracy and greater efficiency during data preparation.

A further barrier in effective data usage is the limited access for non-technical staff, who may struggle with complex database query languages. LMs provide a solution by enabling natural language database interfaces, allowing users to interact with and query datasets in plain language rather than technical commands. This promotes broader participation in data analysis, enabling more stakeholders to benefit from curated information.

Finally, spotting emerging trends and patterns in large datasets has always been a challenge, especially when dealing with dynamic or real-time data. LMs enhance this capability through predictive analytics and pattern detection, identifying correlations, anomalies, and potential developments before they become apparent through traditional analysis. This supports more informed decision-making and empowers organizations to plan strategically and act proactively.

## CONCLUSION

From rigid, rule-based systems to today’s multimodal transformers, LMs have evolved into versatile tools capable of transforming data curation and management. Their ability to automate metadata creation, enable semantic discovery and improve data quality positions them as core assets in modern data ecosystems. As they continue to advance, particularly in reasoning and multimodal integration, LMs will play an even greater role in ensuring that datasets are not only stored but also organized, discoverable, and actionable.

# REFERENCES

Business Insider. (2025, June 8). *Klarna CEO warns AI may cause a recession as the*

*technology comes for white-collar jobs.* [Business Insider](https://www.businessinsider.com/klarna-ceo-ai-may-cause-recession-white-collar-jobs-threat-2025-6?utm_source=chatgpt.com)

Borgman, C. L. (2015). Big Data, Little Data, No Data: *Scholarship in the Networked World*.

Cambridge, MA: MIT Press.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT*: Pre-training of Deep*

*Bidirectional Transformers for Language Understanding*. Proceedings of the 2019

Conference of the North American Chapter of the Association for Computational

Linguistics (NAACL-HLT). Association for Computational Linguistics, Minneapolis,

MN.<https://doi.org/10.18653/v1/N19-1423>

FinTech Weekly. (2025, May 12). *Klarna reverses course on AI customer support, resumes*

*human hiring.* [FinTech Magazine Articles](https://magazine.fintechweekly.com/magazine/articles/klarna-hires-customer-service-after-ai-pivot?utm_source=chatgpt.com)

Forbes. (2025, May 18). *Klarna reverses on AI, says customers like talking to people.*  
 [Forbes](https://www.forbes.com/sites/quickerbettertech/2025/05/18/business-tech-news-klarna-reverses-on-ai-says-customers-like-talking-to-people/?utm_source=chatgpt.com)

Goyal, N., & Ferrara, E. (2023). Large language models and their applications in information

retrieval. *Journal of Data and Information Science*, 8(2), 1–18.

<https://doi.org/10.2478/jdis-2023-0010>

Liang, P. (2022). Advances in natural language interfaces to databases. *Communications of*

*the ACM*, 65(5), 98–107. <https://doi.org/10.1145/3511094>

Min, S., et al. (2023). *FActScore: Fine-grained Atomic Evaluation of Factual Precision in*

*Long Form Text Generation.* [arXiv](https://arxiv.org/abs/2305.14251?utm_source=chatgpt.com)

Tzitzikas, Y., Marketakis, Y., & Minadakis, N. (2021). Semantic search and metadata

enrichment with AI. *International Journal on Digital Libraries*, 22(1), 1–18.

<https://doi.org/10.1007/s00799-020-00302-y>

Vaswani, A., et al. (2017). Attention is all you need. In *Advances in Neural Information*

*Processing Systems* (Vol. 30, pp. 5998–6008). Curran Associates, Inc.

<https://arxiv.org/abs/1706.03762>

Wei, J., et al. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language*

*Models.*