The Evolving Architectures of Large Language Models: A Comprehensive Analysis

1. Introduction

Large Language Models (LLMs) have emerged as a pivotal force in the landscape of artificial intelligence, demonstrating remarkable capabilities in understanding and generating human-like text.1 These sophisticated models have not only redefined the boundaries of natural language processing but have also permeated various domains, showcasing their versatility in tasks ranging from text summarization and translation to code generation and complex reasoning.2 The rapid advancements in LLMs necessitate a thorough understanding of their architectural evolution to grasp the current state and anticipate future trajectories of this dynamic field.1 This report aims to provide a comprehensive analysis of this evolution, tracing the journey from the early foundations of language modeling to the state-of-the-art LLM architectures that underpin today's most advanced AI systems. The scope of this analysis encompasses a historical perspective, highlighting key theoretical developments, methodological innovations, and the shifts in research focus that have shaped the field. Furthermore, this report will delve into the prominent debates surrounding LLMs, analyze citation patterns to identify seminal works and emerging trends, and discuss the reliability and limitations of the existing literature. The methodology employed involves a systematic review of academic sources and an analysis of their citation patterns to provide a holistic view of the evolution of LLM architectures.

Evolution of Large Language Models: A Timeline

- **1950s-1970s:** The earliest stages of AI research focused on rule-based systems for natural language processing. A notable example was ELIZA (1966), one of the first chatbots, which simulated conversation using predefined rules.
- **1980s-1990s:** This period saw a shift towards statistical language models (SLMs) that learned patterns from large text datasets. N-gram models became popular for predicting word sequences based on their frequency.
- **1997:** Long Short-Term Memory (LSTM) networks were developed by Hochreiter and Schmidhuber. LSTMs addressed the vanishing gradient problem in recurrent neural networks (RNNs), enabling the capture of long-term dependencies in language.
- **2010s:** The rise of neural networks and deep learning led to significant advancements in language modeling. Word embeddings, such as Word2Vec (2013), improved the ability of AI to understand semantic relationships between words.
- **2014:** Sequence-to-Sequence (Seq2Seq) models and attention mechanisms gained prominence, particularly for tasks like machine translation. The attention mechanism enhanced the performance of neural machine translation systems.
- **2017:** The Transformer architecture was introduced in the paper "Attention is All You

Need" by Vaswani et al. This architecture revolutionized the field by using attention mechanisms instead of recurrence, allowing for parallel processing and better capture of long-range dependencies.
- **2018:** BERT (Bidirectional Encoder Representations from Transformers) was introduced by Google, improving context understanding through bidirectional training. OpenAI launched GPT-1, demonstrating the power of the Transformer architecture for language understanding via unsupervised pre-training.
- **2019:** OpenAI released GPT-2, a larger version of GPT-1 with 1.5 billion parameters, showcasing improved text generation capabilities. RoBERTa (A Robustly Optimized BERT Pretraining Approach) by Facebook refined BERT's pre-training, achieving state-of-the-art results.
- **2020:** GPT-3, with 175 billion parameters, was released by OpenAI, demonstrating strong zero-shot and few-shot learning abilities. Google introduced ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately), a more sample-efficient pre-training method.
- **2022:** Google introduced PaLM (Pathways Language Model), a massive 540 billion parameter model, highlighting the benefits of scaling. OpenAI's ChatGPT, based on the GPT series, gained widespread public attention for its conversational abilities.
- **2023:** OpenAI released GPT-4, an even more powerful and versatile model with multimodality. Meta AI introduced LLaMA, a family of open and efficient foundation language models, making advanced LLMs more accessible.
- **2024:** Continued advancements in LLMs included improved reasoning, longer context windows, and multimodality (processing text, images, audio, etc.).

2. The Genesis of Language Models: From Rules to Statistics

The quest to enable machines to understand and process human language dates back to the early days of artificial intelligence. Initial attempts in the 1950s and 1960s focused on rule-based systems, where linguistic rules were manually encoded to parse and generate text. A notable example from this era is ELIZA, one of the first chatbots, which simulated conversation by matching user input to pre-programmed responses.4 While these rule-based systems represented a foundational step, they struggled to handle the inherent complexity and variability of natural language.1 The vast nuances, ambiguities, and context-dependent meanings of human language proved difficult to capture through rigid sets of predefined rules, limiting the real-world applicability of these early systems.14

The late 20th century witnessed a significant paradigm shift in language modeling approaches with the advent of statistical language models (SLMs) in the 1980s and 1990s.[1] Instead of relying solely on manually encoded rules, statistical models leveraged probabilistic methods to learn patterns from large text corpora.[1] A key innovation during this period was the development of n-gram models, which became popular for modeling the probability of word sequences based on their frequency of occurrence.[1] These models introduced a mechanism for understanding context in language by focusing on the local relationships between words.[20] The transition to statistical methods marked a fundamental change in how machines processed language, moving away from manual encoding towards learning patterns and structures directly from data.[20] This approach paved the way for more nuanced language analysis and set the stage for the subsequent rise of neural network-based language models.

3. The Neural Network Era: Embracing Deep Learning

The early 2010s marked a turning point in the history of language models with the resurgence of interest in neural networks, particularly deep learning techniques, following their success in image processing around 2012.1 Recurrent Neural Networks (RNNs) emerged as a powerful architecture for processing sequential data, offering the ability to maintain a memory of previous inputs and capture temporal dependencies in language.1
A significant advancement in RNN architecture was the development of Long Short-Term Memory (LSTM) networks in 1997 by Hochreiter and Schmidhuber.[1] LSTMs were designed to address the vanishing gradient problem, a key limitation of traditional RNNs that hindered their ability to learn from long sequences.[45] By incorporating gated units, including input, forget, and output gates, LSTMs could regulate the flow of information through the network over time, enabling them to capture long-term dependencies in sequential data, making them particularly suitable for tasks involving text generation, sentiment analysis, and language modeling.[1] Other variants of RNNs, such as Gated Recurrent Units (GRUs), also emerged, offering similar capabilities with a slightly different architecture.[1] The advent of RNNs, especially LSTMs, allowed neural networks to model the sequential nature of language more effectively, capturing longer-range dependencies that were beyond the capability of statistical models.[26]

Despite their advancements, RNNs faced limitations, particularly when dealing with very long sequences. The sequential processing nature of RNNs made it challenging to parallelize computations, leading to longer training times, especially with large datasets.[1] The vanishing gradient problem also persisted to some extent, making it difficult for RNNs to effectively learn dependencies across very long distances in a sequence.[1] These limitations paved the way for the development of a novel architecture that would revolutionize the field of natural language processing.

4. The Transformer Breakthrough: Attention is All You Need

The year 2017 marked a watershed moment in the evolution of neural network architectures for language processing with the introduction of the Transformer by Vaswani et al. in their seminal paper "Attention Is All You Need".1 This groundbreaking work proposed a novel neural network architecture that eschewed recurrence entirely, relying instead on an attention mechanism to model global dependencies between input and output sequences.30

The Transformer architecture introduced several key components that addressed the limitations of RNNs:

- **Attention Mechanism:** At the heart of the Transformer is the attention mechanism, which allows the model to focus on the most relevant parts of the input sequence when processing or generating output.[2] Unlike traditional models that processed words in isolation or sequentially, attention assigns weights to each word based on its relevance to the current task.[75]
- **Self-Attention:** A key innovation was the self-attention mechanism, which enables the model to attend to different positions of its input sequence to compute a representation of that sequence.[1] This allows the model to weigh the importance of each word in the sequence relative to others, capturing dependencies between different words in the input, regardless of their distance.[1]
- **Multi-Head Attention:** To further enhance the model's ability to capture diverse contextual information, the Transformer employs multi-head attention.[6] This mechanism performs multiple parallel self-attention operations, each with its own set of learned query, key, and value transformations, allowing the model to focus on different aspects of the relationships between words simultaneously.[58]
- **Positional Encoding:** Since the Transformer lacks the inherent sequential processing of RNNs, it uses positional encodings to provide the model with information about the position of each token within the sequence.[6] These encodings are added to the input embeddings, allowing the model to understand the order of words in the sentence.[61]

The Transformer architecture revolutionized the field by replacing recurrent connections with attention mechanisms, which allowed for parallel processing of the entire input sequence, significantly speeding up training and inference.[6] Moreover, the attention mechanism enabled the model to effectively capture long-range dependencies in text, overcoming a key limitation of RNNs.[6] The success of the Transformer architecture laid the foundation for the development of a new generation of powerful language models.

5. Key LLM Architectures and Their Evolution

5.1 The GPT Family: Generative Pre-trained Transformers

The Generative Pre-trained Transformer (GPT) series, pioneered by OpenAI, represents a significant leap in the evolution of LLMs, showcasing the power of scaling Transformer architectures for generative tasks.

- **GPT-1 (2018):** Introduced in 2018, GPT-1 marked an early success in leveraging the Transformer architecture for language understanding through unsupervised pre-training.[4] Utilizing a 12-layer decoder-only Transformer with masked self-attention heads, GPT-1 was pre-trained on the BookCorpus dataset, which contained over 7,000 unpublished fiction books, chosen for its long passages of continuous text that helped the model learn to handle long-range information.[110] This initial model, with 117 million parameters, demonstrated strong performance on various natural language processing tasks after fine-tuning, outperforming discriminatively-trained models on tasks like natural language inference, question answering, and semantic similarity.[110] GPT-1 established the paradigm of generative pre-training followed by task-specific fine-tuning, which became a standard procedure in NLP.[111]
- **GPT-2 (2019):** Released in 2019, GPT-2 was conceived as a "direct scale-up" of GPT-1, featuring a ten-fold increase in both its parameter count (1.5 billion) and the size of its training dataset, which comprised 8 million web pages.[4] Maintaining the decoder-only Transformer architecture, GPT-2 exhibited remarkable abilities in generating coherent and contextually relevant text over extended passages, even demonstrating zero-shot capabilities in tasks like translation, question answering, and summarization.[123] The sheer scale of GPT-2 underscored the benefits of increasing model capacity, leading to significant advancements in language generation.
- **GPT-3 (2020):** Introduced in 2020, GPT-3 marked another substantial leap in scale, boasting 175 billion parameters, an order of magnitude larger than its predecessor.[4] Trained on a vast and diverse dataset of text and code, GPT-3 demonstrated strong zero-shot and few-shot learning abilities across a wide range of NLP tasks without requiring task-specific fine-tuning.[28] Its capabilities extended beyond mere text generation to include translation, question answering, and even code generation, highlighting the emergent abilities that arise from scaling language models to unprecedented sizes.[28]
- **Beyond GPT-3:** The GPT series has continued to evolve with models like GPT-4 [4] and GPT-4V [155], which feature even larger parameter counts and enhanced capabilities, including multimodality. These advancements underscore the ongoing trend of scaling and refining the decoder-only Transformer architecture for increasingly sophisticated language processing and generation.

The GPT family's focus on the decoder-only Transformer architecture has proven particularly effective for generative tasks, where the model predicts the next token given the preceding context.[18] The ability of GPT models to perform zero-shot, one-shot, and few-shot learning has revolutionized how language models are applied, suggesting that these models acquire a broad understanding of language and the world during pre-training, which can be leveraged for diverse downstream tasks through simple prompting.[28]

5.2 BERT and its Variants: Bidirectional Encoder Representations from Transformers

In contrast to the GPT series' focus on generation, Bidirectional Encoder Representations from Transformers (BERT), introduced by Google in 2018, revolutionized the field by focusing on natural language understanding.1
BERT's key innovation was its bidirectional encoder approach, which allowed the model to consider both the left and right context of a word in a sentence, leading to a deeper understanding of language nuances.[166] BERT was pre-trained using two main tasks: Masked Language Modeling (MLM), where the model is trained to predict randomly masked words in a sentence, and Next Sentence Prediction (NSP), where the model learns to understand the relationship between pairs of sentences.[165] This pre-training strategy enabled BERT to learn contextual, latent representations of tokens, making it highly effective for a wide range of natural language understanding tasks, such as question answering, sentiment analysis, and named entity recognition.[1]

Building on the success of BERT, several improvements and variants were developed:

- **RoBERTa (2019):** A Robustly Optimized BERT Pretraining Approach, RoBERTa, introduced in 2019, refined the pre-training procedure of BERT.[2] By training the model for a longer duration with larger batch sizes and more data, and by removing the Next Sentence Prediction task, RoBERTa achieved state-of-the-art results on various natural language understanding benchmarks, often surpassing the performance of the original BERT model.[169]
- **ELECTRA (2020):** Efficiently Learning an Encoder that Classifies Token Replacements Accurately, ELECTRA, introduced in 2020, proposed a more sample-efficient pre-training task called Replaced Token Detection (RTD).[28] Instead of masking tokens, ELECTRA corrupts the input by replacing some tokens with plausible alternatives sampled from a small generator network. A discriminator model is then trained to predict whether each token in the corrupted input was an original or a replaced token.[194] This discriminative pre-training task proved to be more efficient than BERT's generative MLM task, allowing ELECTRA to achieve strong results with less compute.[194]

BERT and its variants highlighted the effectiveness of bidirectional training for natural language understanding, establishing a new standard for pre-trained language models that could be fine-tuned for a wide array of downstream tasks with remarkable success.

5.3 The T5 Model: A Unified Text-to-Text Transformer

The Text-to-Text Transfer Transformer (T5), introduced by Google in 2019, presented a paradigm shift by proposing a unified framework where all natural language processing tasks are treated as text-to-text problems.28
T5 utilizes a standard Transformer architecture with both an encoder and a decoder.[82] The key innovation lies in its approach of framing every NLP task, including translation, question answering, and classification, as a text generation task. This is achieved by feeding the model text as input and training it to generate some target text.[160] To instruct the model on the specific task, the input text is prepended with a task-specific prefix, such as "translate English to German:" or "summarize:".[163] This unified text-to-text framework allowed T5 to handle various tasks using the same model, loss function, hyperparameters, and training procedure, simplifying the process of transfer learning and reducing the complexity of developing separate models for each task.[160]

Variants of T5, such as Flan-T5, further explored the benefits of instruction tuning, demonstrating improved performance on a wide range of tasks by fine-tuning the model on a collection of instances formatted as natural language instructions, inputs, and desired outputs.[28] The T5 model's unified approach showcased the power of a versatile architecture capable of addressing diverse NLP challenges through a consistent text generation paradigm.

5.4 PaLM and Pathways: Scaling Language Modeling

The Pathways Language Model (PaLM), introduced by Google in 2022, represented a significant push towards scaling language models to unprecedented sizes, leveraging the novel Pathways system for highly efficient training across thousands of accelerator chips.2
PaLM is a densely activated, autoregressive Transformer model with 540 billion parameters, trained on 780 billion tokens.[140] It utilizes a decoder-only Transformer architecture with several modifications aimed at improving training efficiency and model performance, including SwiGLU activations, parallel layers with residual connections, multi-query attention, RoPE embeddings, and shared input-output embeddings.[208] A key finding of the PaLM research was the continued benefits of scaling, with the 540 billion parameter model achieving state-of-the-art few-shot

learning results on hundreds of language understanding and generation benchmarks, even outperforming fine-tuned state-of-the-art models on multi-step reasoning tasks and surpassing average human performance on the BIG-bench benchmark.[208] PaLM also demonstrated strong capabilities in multilingual tasks and source code generation.[208] Furthermore, the PaLM family includes PaLM-E, an embodied multimodal language model, showcasing the ability to integrate language understanding with other modalities.[2] The scale and performance of PaLM underscored the potential of extremely large language models and the importance of efficient training systems like Pathways.

5.5 LLaMA and the Rise of Open-Source Models

The introduction of LLaMA (Large Language Model Meta AI) by Meta AI in February 2023 marked a significant shift towards open and efficient foundation language models.2 LLaMA is a collection of foundation language models ranging from 7 billion to 65 billion parameters, trained on trillions of tokens using publicly available datasets exclusively, without relying on proprietary data.141

Like GPT-3, the LLaMA series employs an autoregressive decoder-only Transformer architecture with minor differences such as the use of SwiGLU activation function, rotary positional embeddings (RoPE), and RMSNorm.[141] Despite its smaller size compared to models like GPT-3 and PaLM, LLaMA demonstrated competitive performance on most benchmarks, with LLaMA-13B even outperforming GPT-3 (175B) on many tasks.[141] The release of LLaMA's inference code under an open-source license democratized access to state-of-the-art LLMs, fostering a surge of research and development within the open-source community.[141] Subsequent versions, including LLaMA 2 and LLaMA 3, have continued to build upon this foundation, increasing the model sizes, training data, and capabilities, further solidifying LLaMA as a leading open-weight LLM.[141] The LLaMA family's commitment to open access has significantly accelerated the progress and accessibility of large language model research and applications.

6. Shifts in Research Focus and Methodological Innovations

The evolution of LLM architectures has been accompanied by significant shifts in research focus and the development of innovative methodologies.

Pre-training techniques have evolved considerably from the initial language modeling objectives. BERT introduced masked language modeling and next sentence prediction [165], while T5 unified various tasks under a text-to-text framework.[160] More recent models have explored variations and optimizations of these pre-training objectives to improve the quality of learned representations and the efficiency of the pre-training process.[28]

Transfer learning has become a cornerstone in the application of LLMs, where models pre-trained on massive datasets are fine-tuned on smaller, task-specific datasets to achieve state-of-the-art performance.[2] This approach allows researchers and practitioners to leverage the vast knowledge acquired by large models during pre-training for a wide range of downstream applications with significantly reduced data and computational requirements for fine-tuning.

More recently, instruction tuning has emerged as a crucial technique for enhancing the ability of LLMs to follow natural language instructions and generalize to new tasks.[1] By fine-tuning LLMs on datasets of instructions paired with desired outputs, these models learn to better understand and execute a wide variety of commands, leading to improved task generalization and performance. Reinforcement learning from human feedback (RLHF) has also become a key methodology for aligning LLM behavior with human values and preferences, such as helpfulness, honesty, and harmlessness.[1] By training models to optimize for human-generated rewards based on feedback data, RLHF helps to ensure that LLMs produce outputs that are more aligned with human expectations and ethical considerations.

Another significant trend in the field is the move towards multimodality. LLMs are increasingly being developed with the capability to process and generate not only text but also other data types such as images, audio, and video.[2] This advancement enables LLMs to tackle a broader range of real-world applications by understanding and reasoning across different modalities.

Finally, to address the computational challenges of training and deploying increasingly large models, researchers have explored Mixture of Experts (MoE) architectures.[1] MoE models feature a sparsely-activated expert layer, where different parts of the network are activated for different inputs, allowing for a significant increase in the number of parameters while maintaining a manageable computational cost per example.[217] This approach has enabled the development of models with trillions of parameters, pushing the boundaries of what LLMs can achieve.

7. Prominent Debates and Controversies

The rapid progress in Large Language Models has also sparked several prominent debates and controversies within the AI community and beyond.
One significant area of discussion revolves around the interpretability and explainability of LLMs.[44] Often characterized as "black box" models, the intricate

workings and decision-making processes of LLMs can be opaque, raising concerns about transparency and accountability, especially when applied in critical domains.[223] Understanding why and how these models arrive at specific conclusions remains a challenge, hindering the ability to fully trust and debug their outputs.[311]

Another major controversy surrounds the issues of bias and fairness in LLMs.[1] Trained on vast amounts of uncurated internet data, LLMs can inherit and even amplify harmful social biases present in their training data, leading to outputs that are discriminatory, stereotypical, or misrepresentative of certain demographic groups.[1] The potential for misuse of LLMs, including the generation of misinformation, disinformation, and toxic content, also raises significant ethical concerns.[250]

The computational costs and environmental impact associated with training and deploying very large language models have also been a subject of debate.[4] The immense scale of these models, often involving billions or even trillions of parameters, requires substantial computational resources and energy consumption, raising questions about sustainability and accessibility.[125]

Finally, an ongoing debate persists about whether LLMs truly understand language and possess genuine intelligence.[1]

8. Analysis of Citation Patterns

The analysis of citation patterns within the literature on LLM architectures reveals several key insights into the evolution and impact of this field. Seminal works with exceptionally high citation counts highlight the foundational contributions that have shaped the trajectory of LLM development. Notably, the paper "Attention Is All You Need" 77, which introduced the Transformer architecture, stands as a cornerstone, having garnered over 173,000 citations as of 2025.77 This paper's impact is evident in its widespread adoption as the underlying architecture for most modern LLMs.77 Similarly, the paper on Long Short-Term Memory (LSTM) by Hochreiter and Schmidhuber 50, with over 126,000 citations 50, marks a crucial development in enabling RNNs to learn long-range dependencies, paving the way for more sophisticated sequence modeling. The word2vec papers by Mikolov et al. 395, with tens of thousands of citations each, revolutionized the field of word embeddings, providing efficient methods for learning high-quality vector representations of words from large datasets, which are fundamental to many subsequent LLM architectures.

Emerging research trends are discernible through the analysis of more recent publications and their citation patterns. The field is currently witnessing significant interest in efficient Transformer architectures aimed at reducing computational costs and memory footprint.[2] Research on effectively handling longer context lengths in Transformers is also gaining traction, as the ability to process and understand longer sequences is crucial for many real-world applications. Furthermore, the trend towards multimodal LLMs, capable of processing and generating information across various modalities like text, images, and audio, is increasingly prominent in recent research.[2]

Identifying underexplored research areas requires a deeper analysis of citation patterns, looking for works that may have been overlooked or areas where further investigation is needed.[2] For instance, while the scaling of model size has been extensively explored, the theoretical underpinnings of emergent abilities and the optimal strategies for efficient fine-tuning in various domains might warrant further investigation.

Interdisciplinary connections can be identified by examining the diverse range of authors, the journals and conferences where they publish, and the works they cite.[1] The literature spans across

Works cited

1. The Evolution and Impact of Large Language Model Systems: A Comprehensive Analysis, accessed on May 1, 2025, https://www.researchgate.net/publication/379091956_The_Evolution_and_Impact_of_Large_Language_Model_Systems_A_Comprehensive_Analysis
2. (PDF) Survey of Different Large Language Model Architectures: Trends, Benchmarks, and Challenges - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/383976933_Survey_of_different_Large_Language_Model_Architectures_Trends_Benchmarks_and_Challenges
3. [2303.18223] A Survey of Large Language Models - arXiv, accessed on May 1, 2025, https://arxiv.org/abs/2303.18223
4. History and Evolution of LLMs | GeeksforGeeks, accessed on May 1, 2025, https://www.geeksforgeeks.org/history-and-evolution-of-llms/
5. A Review of Large Language Models: Fundamental Architectures, Key Technological Evolutions, Interdisciplinary Technologies Integration, Optimization and Compression Techniques, Applications, and Challenges - MDPI, accessed on May 1, 2025, https://www.mdpi.com/2079-9292/13/24/5040
6. A Historical Survey of Advances in Transformer Architectures - MDPI, accessed on May 1, 2025, https://www.mdpi.com/2076-3417/14/10/4316
7. A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs) - MDPI, accessed on May 1, 2025, https://www.mdpi.com/2076-3417/14/5/2074
8. Large language model - Wikipedia, accessed on May 1, 2025, https://en.wikipedia.org/wiki/Large_language_model
9. Large Language Models: What You Need to Know in 2025 | HatchWorks AI, accessed on May 1, 2025, https://hatchworks.com/blog/gen-ai/large-language-models-guide/
10. Evolution of Large Language Models (2025 Updated) - The Expert Community,

accessed on May 1, 2025, https://theexpertcommunity.com/artificial-intelligence/evolution-of-large-language-models/
11. Timeline of large language models, accessed on May 1, 2025, https://timelines.issarice.com/wiki/Timeline_of_large_language_models
12. Large Language Models 101: History, Evolution and Future - Scribble Data, accessed on May 1, 2025, https://www.scribbledata.io/blog/large-language-models-history-evolutions-and-future/
13. Master NLP History: From Then to Now - Shelf.io, accessed on May 1, 2025, https://shelf.io/blog/master-nlp-history-from-then-to-now/
14. History and Evolution of NLP | GeeksforGeeks, accessed on May 1, 2025, https://www.geeksforgeeks.org/history-and-evolution-of-nlp/
15. History Of Natural Language Processing - Let's Data Science, accessed on May 1, 2025, https://letsdatascience.com/learn/history/history-of-natural-language-processing/
16. Natural language processing - Wikipedia, accessed on May 1, 2025, https://en.wikipedia.org/wiki/Natural_language_processing
17. The History of Natural Language Processing — Leximancer Qualitative Research | Thematic Analysis | Map, accessed on May 1, 2025, https://www.leximancer.com/blog/kxpw5rc8ojnxv8106yr3et22wmn5zi
18. History, Development, and Principles of Large Language Models—An Introductory Survey, accessed on May 1, 2025, https://arxiv.org/html/2402.06853v1
19. Language model - Wikipedia, accessed on May 1, 2025, https://en.wikipedia.org/wiki/Language_model
20. Evolution of Language Models: From Rules-Based Models to LLMs - Appy Pie, accessed on May 1, 2025, https://www.appypie.com/blog/evolution-of-language-models
21. Evolution and Prospects of Foundation Models: From Large Language Models to Large Multimodal Models - SciOpen, accessed on May 1, 2025, https://www.sciopen.com/article/10.32604/cmc.2024.052618
22. 1. History Of Large Language Models | From 1940 To 2023 - AI Researcher, accessed on May 1, 2025, https://ai-researchstudies.com/history-of-large-language-models-from-1940-to-2023/
23. Language Models: Past, Present, and Future - Communications of the ACM, accessed on May 1, 2025, https://cacm.acm.org/research/language-models/
24. A Brief History of Large Language Models - DATAVERSITY, accessed on May 1, 2025, https://www.dataversity.net/a-brief-history-of-large-language-models/
25. Evolution of Neural Networks to Large Language Models - Labellerr, accessed on May 1, 2025, https://www.labellerr.com/blog/evolution-of-neural-networks-to-large-language-models/
26. Gentle Introduction to Statistical Language Modeling and Neural Language

Models - MachineLearningMastery.com, accessed on May 1, 2025, https://machinelearningmastery.com/statistical-language-modeling-and-neural-language-models/
27. History of Evolution of Language Models in NLP: Foundation of Generative AI and LLM, accessed on May 1, 2025, https://www.youtube.com/watch?v=ohdvSwKaQDk
28. A Comprehensive Overview of Large Language Models - arXiv, accessed on May 1, 2025, https://arxiv.org/html/2307.06435v9
29. From statistics to deep learning: Using large language models in psychiatric research - PMC, accessed on May 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11707704/
30. Attention is All you Need - NIPS papers, accessed on May 1, 2025, https://papers.neurips.cc/paper/7181-attention-is-all-you-need.pdf
31. Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications - MDPI, accessed on May 1, 2025, https://www.mdpi.com/2078-2489/15/9/517
32. Andrej Karpathy, accessed on May 1, 2025, https://karpathy.ai/
33. What is RNN? - Recurrent Neural Networks Explained - AWS, accessed on May 1, 2025, https://aws.amazon.com/what-is/recurrent-neural-network/
34. What are the advantages and disadvantages of a Recurrent Neural Network (RNN)?, accessed on May 1, 2025, https://aiml.com/what-are-the-advantages-and-disadvantages-of-a-recurrent-neural-network-rnn/
35. Introduction to Recurrent Neural Networks | GeeksforGeeks, accessed on May 1, 2025, https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/
36. Recurrent Neural Network Language Models, accessed on May 1, 2025, http://phontron.com/class/mtandseq2seq2018/assets/slides/mt-fall2018.chapter6.pdf
37. Recurrent neural network - Wikipedia, accessed on May 1, 2025, https://en.wikipedia.org/wiki/Recurrent_neural_network
38. Sequence Processing with Recurrent Networks - Stanford University, accessed on May 1, 2025, https://web.stanford.edu/~jurafsky/slp3/old_oct19/9.pdf
39. Survey on Recurrent Neural Network in Natural Language Processing - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/319937209_Survey_on_Recurrent_Neural_Network_in_Natural_Language_Processing
40. A Review of the Neural History of Natural Language Processing - ruder.io, accessed on May 1, 2025, https://www.ruder.io/a-review-of-the-recent-history-of-nlp/
41. Natural language processing and recurrent network models for identifying genomic mutation-associated cancer treatment change from patient progress notes - PMC, accessed on May 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC6435007/
42. Chapter 4 Recurrent neural networks and their applications in NLP | Modern Approaches in Natural Language Processing, accessed on May 1, 2025,

https://slds-lmu.github.io/seminar_nlp_ss20/recurrent-neural-networks-and-their-applications-in-nlp.html
43. Where can I find the original paper that introduced RNNs? - AI Stack Exchange, accessed on May 1, 2025, https://ai.stackexchange.com/questions/8190/where-can-i-find-the-original-paper-that-introduced-rnns
44. The Unreasonable Effectiveness of Recurrent Neural Networks - Andrej Karpathy blog, accessed on May 1, 2025, http://karpathy.github.io/2015/05/21/rnn-effectiveness/
45. (PDF) Long Short-Term Memory - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/13853244_Long_Short-Term_Memory
46. LSTM Explained - Papers With Code, accessed on May 1, 2025, https://paperswithcode.com/method/lstm
47. Long short-term memory - Wikipedia, accessed on May 1, 2025, https://en.wikipedia.org/wiki/Long_short-term_memory
48. J. Schmidhuber - Semantic Scholar, accessed on May 1, 2025, https://www.semanticscholar.org/author/J.-Schmidhuber/145341374
49. Long short-term memory - BibSonomy, accessed on May 1, 2025, https://www.bibsonomy.org/bibtex/2a4a80026d24955b267cae636aa8abe4a/dallmann
50. Juergen Schmidhuber - Google Scholar, accessed on May 1, 2025, https://scholar.google.com/citations?user=gLnCTgIAAAAJ&hl=en
51. Long Short-Term Memory - Deep Learning, CMU, accessed on May 1, 2025, https://deeplearning.cs.cmu.edu/S23/document/readings/LSTM.pdf
52. LONG SHORT-TERM MEMORY 1 INTRODUCTION, accessed on May 1, 2025, https://www.bioinf.jku.at/publications/older/2604.pdf
53. The most cited neural networks all build on work done in my labs, accessed on May 1, 2025, https://people.idsia.ch/~juergen/most-cited-neural-nets.html
54. Long Short-Term Memory-Networks for Machine Reading - ACL Anthology, accessed on May 1, 2025, https://aclanthology.org/D16-1053.pdf
55. Transformer (deep learning architecture) - Wikipedia, accessed on May 1, 2025, https://en.wikipedia.org/wiki/Transformer_(deep_learning_architecture)
56. Transformer: A Novel Neural Network Architecture for Language Understanding, accessed on May 1, 2025, https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/
57. Attention (machine learning) - Wikipedia, accessed on May 1, 2025, https://en.wikipedia.org/wiki/Attention_(machine_learning)
58. Self – attention in NLP | GeeksforGeeks, accessed on May 1, 2025, https://www.geeksforgeeks.org/self-attention-in-nlp/
59. Transformer Model: Impact, Architecture, and 5 Types of Transformers - Kolena, accessed on May 1, 2025, https://www.kolena.com/guides/transformer-model-impact-architecture-and-5-types-of-transformers/
60. What is a Transformer Model? - IBM, accessed on May 1, 2025,

https://www.ibm.com/think/topics/transformer-model
61. How Transformers Work: A Detailed Exploration of Transformer Architecture - DataCamp, accessed on May 1, 2025, https://www.datacamp.com/tutorial/how-transformers-work
62. Transformers in Machine Learning | GeeksforGeeks, accessed on May 1, 2025, https://www.geeksforgeeks.org/getting-started-with-transformers/
63. What are Transformers in Artificial Intelligence? - AWS, accessed on May 1, 2025, https://aws.amazon.com/what-is/transformers-in-artificial-intelligence/
64. How do Transformers work? - Hugging Face LLM Course, accessed on May 1, 2025, https://huggingface.co/learn/llm-course/chapter1/4
65. Transformer Architecture - Lark, accessed on May 1, 2025, https://www.larksuite.com/en_us/topics/ai-glossary/transformer-architecture
66. Transformers: a Primer, accessed on May 1, 2025, http://www.columbia.edu/~jsl2239/transformers.html
67. A Deep Dive Into the Transformer Architecture – The Development of Transformer Models | Exxact Blog, accessed on May 1, 2025, https://www.exxactcorp.com/blog/Deep-Learning/a-deep-dive-into-the-transformer-architecture-the-development-of-transformer-models
68. [D] Are recurrent neural networks being phased out? : r/MachineLearning - Reddit, accessed on May 1, 2025, https://www.reddit.com/r/MachineLearning/comments/bkwms8/d_are_recurrent_neural_networks_being_phased_out/
69. Selective State Space Models Outperform Transformers at Predicting RNA-Seq Read Coverage - PMC, accessed on May 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11870438/
70. Applications of transformer-based language models in bioinformatics: a survey - Oxford Academic, accessed on May 1, 2025, https://academic.oup.com/bioinformaticsadvances/article/3/1/vbad001/6984737
71. Transformer Architecture and Attention Mechanisms in Genome Data Analysis: A Comprehensive Review - PMC, accessed on May 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10376273/
72. From Turing to Transformers: A Comprehensive Review and Tutorial on the Evolution and Applications of Generative Transformer Models - MDPI, accessed on May 1, 2025, https://www.mdpi.com/2413-4155/5/4/46
73. CovTransformer: A transformer model for SARS-CoV-2 lineage frequency forecasting, accessed on May 1, 2025, https://www.medrxiv.org/content/10.1101/2024.04.01.24305089v1.full-text
74. Introduction to Large Language Models | Machine Learning - Google for Developers, accessed on May 1, 2025, https://developers.google.com/machine-learning/resources/intro-llms
75. Attention Mechanism in LLMs: An Intuitive Explanation - DataCamp, accessed on May 1, 2025, https://www.datacamp.com/blog/attention-mechanism-in-llms-intuition
76. Attention is all you need: utilizing attention in AI-enabled drug discovery - PubMed Central, accessed on May 1, 2025,

https://pmc.ncbi.nlm.nih.gov/articles/PMC10772984/
77. Attention Is All You Need - Wikipedia, accessed on May 1, 2025, https://en.wikipedia.org/wiki/Attention_Is_All_You_Need
78. Attention is All You Need - Google Research, accessed on May 1, 2025, https://research.google/pubs/attention-is-all-you-need/
79. Leave No Context Behind: Efficient Infinite Context Transformers with Infini-attention - arXiv, accessed on May 1, 2025, https://arxiv.org/html/2404.07143v1
80. What exactly are keys, queries, and values in attention mechanisms? - Stats Stackexchange, accessed on May 1, 2025, https://stats.stackexchange.com/questions/421935/what-exactly-are-keys-queries-and-values-in-attention-mechanisms
81. [1706.03762] Attention Is All You Need - arXiv, accessed on May 1, 2025, https://arxiv.org/abs/1706.03762
82. A Unified Text-to-Text Framework for NLP Tasks: An Overview of T5 Model, accessed on May 1, 2025, https://blog.paperspace.com/flan-t5-architecture/
83. Comparison Between BERT and GPT-3 Architectures | Baeldung on Computer Science, accessed on May 1, 2025, https://www.baeldung.com/cs/bert-vs-gpt-3-architecture
84. Decoder vs Encoder-decoder clarification - Beginners - Hugging Face Forums, accessed on May 1, 2025, https://discuss.huggingface.co/t/decoder-vs-encoder-decoder-clarification/44330
85. ChatGPT's Architecture - Decoder Only? Or Encoder-Decoder?, accessed on May 1, 2025, https://datascience.stackexchange.com/questions/118260/chatgpts-architecture-decoder-only-or-encoder-decoder
86. Why use Encoder-Decoder Models? - DeepLearning.AI, accessed on May 1, 2025, https://community.deeplearning.ai/t/why-use-encoder-decoder-models/563489
87. [D]Encoder only vs encoder-decoder vs decoder only : r/MachineLearning - Reddit, accessed on May 1, 2025, https://www.reddit.com/r/MachineLearning/comments/14y7ajc/dencoder_only_vs_encoderdecoder_vs_decoder_only/
88. Which situation will helpful using encoder or decoder or both in transformer model?, accessed on May 1, 2025, https://ai.stackexchange.com/questions/41505/which-situation-will-helpful-using-encoder-or-decoder-or-both-in-transformer-mod
89. Attention Mechanisms in NLP – Let's Understand the What and Why - Wissen, accessed on May 1, 2025, https://www.wissen.com/blog/attention-mechanisms-in-nlp---lets-understand-the-what-and-why
90. What is an attention mechanism? - IBM, accessed on May 1, 2025, https://www.ibm.com/think/topics/attention-mechanism
91. Revolutionary Attention Mechanism: Power of Transformers - Data Science Dojo, accessed on May 1, 2025,

https://datasciencedojo.com/blog/understanding-attention-mechanism/
92. Understanding and Coding the Self-Attention Mechanism of Large Language Models From Scratch - Sebastian Raschka, accessed on May 1, 2025, https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html
93. Chapter 8 Attention and Self-Attention for NLP | Modern Approaches in Natural Language Processing, accessed on May 1, 2025, https://slds-lmu.github.io/seminar_nlp_ss20/attention-and-self-attention-for-nlp.html
94. Attention mechanism: Overview - YouTube, accessed on May 1, 2025, https://www.youtube.com/watch?v=fjJOgb-E41w
95. [D] How to truly understand attention mechanism in transformers? : r/MachineLearning - Reddit, accessed on May 1, 2025, https://www.reddit.com/r/MachineLearning/comments/qidpqx/d_how_to_truly_understand_attention_mechanism_in/
96. But Google is the author of 'Attention Is All You Need', they know how to build LLM's - Reddit, accessed on May 1, 2025, https://www.reddit.com/r/ChatGPT/comments/125i04d/but_google_is_the_author_of_attention_is_all_you/
97. (PDF) Attention is All you Need (2017) | Chat PDF - Nanonets, accessed on May 1, 2025, https://nanonets.com/chat-pdf/attention-is-all-you-need
98. The background needed to understand "Attention is All You Need" Paper - Reddit, accessed on May 1, 2025, https://www.reddit.com/r/learnmachinelearning/comments/17ywtkd/the_background_needed_to_understand_attention_is/
99. Transformer models: an introduction and catalog - arXiv, accessed on May 1, 2025, https://arxiv.org/html/2302.07730v4
100. The Transformer, accessed on May 1, 2025, https://web.stanford.edu/~jurafsky/slp3/9.pdf
101. (Open Access) Attention Is All You Need (2017) | Ashish Vaswani | 7227 Citations, accessed on May 1, 2025, https://scispace.com/papers/attention-is-all-you-need-1hpncqdg1c
102. Attention Is All You Need | Request PDF - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/317558625_Attention_Is_All_You_Need
103. Ashish Vaswani - Google Scholar, accessed on May 1, 2025, https://scholar.google.com/citations?user=oR9sCGYAAAAJ&hl=en
104. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017) Attention Is All You Need. arXiv 1706.03762. - References, accessed on May 1, 2025, https://www.scirp.org/reference/referencespapers?referenceid=3700044
105. Attention Is All You Need. - dblp, accessed on May 1, 2025, https://dblp.org/rec/journals/corr/VaswaniSPUJGKP17
106. Attention is All you Need - NIPS papers, accessed on May 1, 2025, https://papers.nips.cc/paper/7181-attention-is-all-you-need
107. Attention is All you Need - Semantic Scholar, accessed on May 1, 2025, https://www.semanticscholar.org/paper/Attention-is-All-you-Need-Vaswani-Shaz

eer/204e3073870fae3d05bcbc2f6a8e263d9b72e776
108. OpenAI GPT-3: Everything You Need to Know [Updated] - Springboard, accessed on May 1, 2025, https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/
109. akshat0123/GPT-1: Pytorch implementation of GPT-1 - GitHub, accessed on May 1, 2025, https://github.com/akshat0123/GPT-1
110. GPT-1 - Wikipedia, accessed on May 1, 2025, https://en.wikipedia.org/wiki/GPT-1
111. Understanding the Evolution of ChatGPT: Part 1-An In-Depth Look at GPT-1 and What Inspired It | Towards Data Science, accessed on May 1, 2025, https://towardsdatascience.com/understanding-the-evolution-of-gpt-part-1-an-in-depth-look-at-gpt-1-and-what-inspired-it-b7388a32e87d/
112. GPT-1 Paper Explained - YouTube, accessed on May 1, 2025, https://m.youtube.com/watch?v=m9UCVqd5lGY&pp=ygUFI2dwdDE%3D
113. 2018 ImprovingLanguageUnderstandingb - GM-RKB, accessed on May 1, 2025, https://www.gabormelli.com/RKB/2018_ImprovingLanguageUnderstandingb
114. Karthik Narasimhan - Google Scholar, accessed on May 1, 2025, https://scholar.google.com/citations?user=euc0GX4AAAAJ&hl=en
115. Improving language understanding by generative pre-training - BibBase, accessed on May 1, 2025, https://bibbase.org/network/publication/radford-narasimhan-salimans-sutskever-improvinglanguageunderstandingbygenerativepretraining-2018
116. Alec Radford - Google Scholar, accessed on May 1, 2025, https://scholar.google.com/citations?user=dOad5HoAAAAJ&hl=en
117. Improving Language Understanding by Generative Pre-Training - OpenAI, accessed on May 1, 2025, https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
118. Improving Language Understanding by Generative Pre-Training | Papers With Code, accessed on May 1, 2025, https://paperswithcode.com/paper/improving-language-understanding-by
119. Improving language understanding by generative pre-training - BibSonomy, accessed on May 1, 2025, https://www.bibsonomy.org/bibtex/5c343ed9a31ac52fd17a898f72af228f
120. JAKET: Joint Pre-training of Knowledge Graph and Language Understanding, accessed on May 1, 2025, https://ojs.aaai.org/index.php/AAAI/article/view/21417/21166
121. Improving language understanding with unsupervised learning - OpenAI, accessed on May 1, 2025, https://openai.com/index/language-unsupervised/
122. Improving Language Understanding by Generative Pre-Training - Semantic Scholar, accessed on May 1, 2025, https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a50

35
123. GPT-3 - Wikipedia, accessed on May 1, 2025, https://en.wikipedia.org/wiki/GPT-3
124. GPT-3: All you need to know about AI language model - Sigmoid, accessed on May 1, 2025, https://www.sigmoid.com/blogs/gpt-3-all-you-need-to-know-about-the-ai-language-model/
125. OpenAI's GPT-3 Language Model: A Technical Overview - Lambda, accessed on May 1, 2025, https://lambdalabs.com/blog/demystifying-gpt-3
126. GPT-2 - Wikipedia, accessed on May 1, 2025, https://en.wikipedia.org/wiki/GPT-2
127. OpenAI GPT2 - Hugging Face, accessed on May 1, 2025, https://huggingface.co/docs/transformers/model_doc/gpt2
128. openai/gpt-2: Code for the paper "Language Models are Unsupervised Multitask Learners", accessed on May 1, 2025, https://github.com/openai/gpt-2
129. GPT-2 Explained | Papers With Code, accessed on May 1, 2025, https://paperswithcode.com/method/gpt-2
130. gpt - Unity, accessed on May 1, 2025, https://docs.unity.rc.umass.edu/documentation/datasets/ai/gpt/
131. Instruction Pre-Training: Language Models are Supervised Multitask Learners - arXiv, accessed on May 1, 2025, https://arxiv.org/html/2406.14491v1
132. Language Models are Unsupervised Multitask Learners - Papers With Code, accessed on May 1, 2025, https://paperswithcode.com/paper/language-models-are-unsupervised-multitask
133. Language Models are Unsupervised Multitask Learners annotated/explained version., accessed on May 1, 2025, https://fermatslibrary.com/s/language-models-are-unsupervised-multitask-learners
134. Language Models are Unsupervised Multitask Learners | OpenAI, accessed on May 1, 2025, https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
135. Language Models are Unsupervised Multitask Learners | BibSonomy, accessed on May 1, 2025, https://www.bibsonomy.org/bibtex/1ce8168300081d74707849ed488e2a458
136. Language Models are Unsupervised Multitask Learners | BibSonomy, accessed on May 1, 2025, https://www.bibsonomy.org/bibtex/1b926ece39c03cdf5499f6540cf63babd
137. Language Models are Unsupervised Multitask Learners - Semantic Scholar, accessed on May 1, 2025, https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe
138. GPT-3 Explained | Papers With Code, accessed on May 1, 2025, https://paperswithcode.com/method/gpt-3

139. GPT-3 Research Assistant | Answer questions with research - Community, accessed on May 1, 2025, https://community.openai.com/t/gpt-3-research-assistant-answer-questions-with-research/14616
140. Papers with Code - PaLM Explained, accessed on May 1, 2025, https://paperswithcode.com/method/palm
141. Llama (language model) - Wikipedia, accessed on May 1, 2025, https://en.wikipedia.org/wiki/Llama_(language_model)
142. LLaMA: Open and Efficient Foundation Language Models | Research - AI at Meta, accessed on May 1, 2025, https://ai.meta.com/research/publications/llama-open-and-efficient-foundation-language-models/
143. LLaMA: Open and Efficient Foundation Language Models - Meta Research - Facebook, accessed on May 1, 2025, https://research.facebook.com/publications/llama-open-and-efficient-foundation-language-models/
144. Language Models are Few-Shot Learners | Connected Papers Search, accessed on May 1, 2025, https://www.connectedpapers.com/search?q=Language%20Models%20are%20Few-Shot%20Learners&p=3
145. GPT-3: Language Models are Few-shot Learners - YouTube, accessed on May 1, 2025, https://www.youtube.com/watch?v=5i-SC-roENM
146. GPT-3: Language Models are Few-Shot Learners (Paper Explained) - YouTube, accessed on May 1, 2025, https://www.youtube.com/watch?v=SY5PvZrJhLE
147. [2005.14165] Language Models are Few-Shot Learners - arXiv, accessed on May 1, 2025, https://arxiv.org/abs/2005.14165
148. Tom B Brown - Google Scholar, accessed on May 1, 2025, https://scholar.google.com/citations?user=RLvsC94AAAAJ&hl=en
149. Language models are few-shot learners - BibSonomy, accessed on May 1, 2025, https://www.bibsonomy.org/bibtex/27a2a9aee490ff30dd5b4d0470a8be8d8/albinzehe
150. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners, accessed on May 1, 2025, https://aclanthology.org/2021.naacl-main.185/
151. Language Models are Few-Shot Learners - NIPS papers, accessed on May 1, 2025, https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
152. Language Models are Few-Shot Learners - arXiv, accessed on May 1, 2025, https://arxiv.org/pdf/2005.14165
153. Language Models are Few-Shot Learners - NIPS papers, accessed on May 1, 2025, https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
154. A Comprehensive Overview of Large Language Models - arXiv, accessed on

May 1, 2025, http://arxiv.org/pdf/2307.06435
155. survey on multimodal large language models | National Science Review - Oxford Academic, accessed on May 1, 2025, https://academic.oup.com/nsr/article/11/12/nwae403/7896414
156. The (R)Evolution of Multimodal Large Language Models: A Survey - Lorenzo Baraldi, accessed on May 1, 2025, https://www.lorenzobaraldi.com/media/news/2024_Multimodal_LLMs_Survey__arXiv_.pdf
157. Large Language Models: A Comprehensive Survey on Architectures, Applications, and Challenges - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/387305663_Large_Language_Models_A_Comprehensive_Survey_on_Architectures_Applications_and_Challenges
158. LLMs are Also Effective Embedding Models: An In-depth Overview - arXiv, accessed on May 1, 2025, https://arxiv.org/html/2412.12591v1
159. Revisiting Word Embeddings in the LLM Era - arXiv, accessed on May 1, 2025, https://arxiv.org/html/2402.11094v3
160. T5 Explained - Papers With Code, accessed on May 1, 2025, https://paperswithcode.com/method/t5
161. What is the T5-Model? - Data Basecamp, accessed on May 1, 2025, https://databasecamp.de/en/ml-blog/t5-model
162. T5 (language model) - Wikipedia, accessed on May 1, 2025, https://en.wikipedia.org/wiki/T5_(language_model)
163. T5 - Hugging Face, accessed on May 1, 2025, https://huggingface.co/docs/transformers/model_doc/t5
164. The T5 Model, accessed on May 1, 2025, https://dkharazi.github.io/notes/ml/nlp/t5/
165. BERT (language model) - Wikipedia, accessed on May 1, 2025, https://en.wikipedia.org/wiki/BERT_(language_model)
166. BERT Explained - Papers With Code, accessed on May 1, 2025, https://paperswithcode.com/method/bert
167. Paper Dissected: "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" Explained - DataScienceToday, accessed on May 1, 2025, https://datasciencetoday.net/index.php/en-us/nlp/211-paper-dissected-bert-pre-training-of-deep-bidirectional-transformers-for-language-understanding-explained
168. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, accessed on May 1, 2025, https://aclanthology.org/N19-1423/
169. (PDF) RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019) | Yinhan Liu | 25520 Citations - SciSpace, accessed on May 1, 2025, https://scispace.com/papers/roberta-a-robustly-optimized-bert-pretraining-approach-2rj0bdyhim
170. RoBERTa: A Robustly Optimized BERT Pretraining Approach - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/334735779_RoBERTa_A_Robustly_Opti

mized_BERT_Pretraining_Approach
171. RoBERTa: A Robustly Optimized BERT Pretraining Approach | OpenReview, accessed on May 1, 2025, https://openreview.net/forum?id=SyxS0T4tvS
172. BERT: Pre-training of deep bidirectional transformers for language understanding - EndNote Web, accessed on May 1, 2025, https://web.endnote.com/citations/eyJkaXNwbGF5VGV4dCI6IihEZXZsaW4gZXQgYWwuKSIsImNpdGF0aW9ucyI6W3siYmlibGlvQ29udGVudCI6W3siZGF0ZSI6IjIwMTkvMTAvLyIsInB1Ymxpc2hlciI6IkFzc29jaWF0aW9uIGZvciBDb21wdXRhdGlvbmFsIExpbmd1aXN0aWNzIiwiZ3VpZCI6IjJjZTA3N2VmLTk2ODctNDRlNi05NmRhLTNkODEyYmRlNDNkNyIsInBsYWNlUHVibGlzaGVkIjoiTWlubmVhcG9saXMsIE1pbm5lc290YSIsInJlZmVyZW5jZVR5cGUiOiIxMCIsImdyb3VwR3VpZHMiOltdLCJyZWNvcmRTdGF0dXMiOiJhY3RpdmUiLCJwYWdlcyI6IjQxNzEtNDE4NiIsImVsZWN0cm9uaWNSZXNvdXJjZU51bWJlciI6IjEwLjE4NjUzL3YxL04xOS0xNDIzIiwidGl0bGUiOiJCRVJUOiBQcmUtdHJhaW5pbmcgb2YgZGVlcCBiaWRpcmVjdGlvbmFsIHRyYW5zZm9ybWVycyBmb3IgbGFuZ3VhZ2UgdW5kZXJzdGFuZGluZyIsImF1dGhvcnMiOlsiRGV2bGluLCBKYWNvYiIsIkNoYW5nLCBNaW5nLVdlaSIsIkxlZSwgS2VudG9uIiwiVG91dGFub3ZhLCBLcmlzdGluYSJdLCJ2b2x1bWUiOiIxIn1dLCJndWlkIjoiMmNlMDc3ZWYtOTY4Ny00NGU2LTk2ZGEtM2Q4MTJiZGU0M2Q3IiwiZ3JvdXBHdWlkcyI6W119XX0%3D
173. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding | Request PDF - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/328230984_BERT_Pre-training_of_Deep_Bidirectional_Transformers_for_Language_Understanding
174. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) BERT Pre-Training of Deep ... - Scientific Research Publishing, accessed on May 1, 2025, https://www.scirp.org/reference/referencespapers?referenceid=2624987
175. [1810.04805] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding - arXiv, accessed on May 1, 2025, https://arxiv.org/abs/1810.04805
176. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. - DBLP, accessed on May 1, 2025, https://dblp.org/rec/conf/naacl/DevlinCLT19
177. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding, accessed on May 1, 2025, https://connections-qj.org/article/bert-pretraining-deep-bidirectional-transformers-language-understanding
178. STAT946F20/BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding - statwiki - Math Wiki Server, accessed on May 1, 2025, https://wiki.math.uwaterloo.ca/statwiki/index.php?title=STAT946F20/BERT:_Pre-training_of_Deep_Bidirectional_Transformers_for_Language_Understanding
179. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, accessed on May 1, 2025, https://www.bibsonomy.org/bibtex/210c860e3f390c6fbfd78a3b91ab9b0af/albinzehe
180. BERT: Pre-training of Deep Bidirectional Transformers for Language

Understanding - Semantic Scholar, accessed on May 1, 2025, https://www.semanticscholar.org/paper/BERT%3A-Pre-training-of-Deep-Bidirectional-for-Devlin-Chang/df2b0e26d0599ce3e70df8a9da02e51594e0e992
181. (PDF) RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019) | Yinhan Liu | 15576 Citations - SciSpace, accessed on May 1, 2025, https://typeset.io/papers/roberta-a-robustly-optimized-bert-pretraining-approach-2rj0bdyhim?citations_page=58
182. RoBERTa: A Robustly Optimized BERT Pretraining Approach - ProQuest, accessed on May 1, 2025, https://www.proquest.com/docview/2266196859?pq-origsite=primo
183. RoBERTa: A Robustly Optimized BERT Pretraining Approach. - DBLP, accessed on May 1, 2025, https://dblp.org/rec/journals/corr/abs-1907-11692
184. [1907.11692] RoBERTa: A Robustly Optimized BERT Pretraining Approach - arXiv, accessed on May 1, 2025, https://arxiv.org/abs/1907.11692
185. Liu, Y., Ott, M., Goyal, N., et al. (2020) RoBERTa A Robustly Optimized BERT Pretraining Approach. - References - Scientific Research Publishing, accessed on May 1, 2025, https://www.scirp.org/reference/referencespapers?referenceid=3023380
186. RoBERTa: A Robustly Optimized BERT Pretraining Approach | Papers With Code, accessed on May 1, 2025, https://paperswithcode.com/paper/roberta-a-robustly-optimized-bert-pretraining
187. [R][1907.11692] RoBERTa: A Robustly Optimized BERT Pretraining Approach - Reddit, accessed on May 1, 2025, https://www.reddit.com/r/MachineLearning/comments/cjbcxm/r190711692_roberta_a_robustly_optimized_bert/
188. RoBERTa: A Robustly Optimized BERT Pretraining Approach - Semantic Scholar, accessed on May 1, 2025, https://www.semanticscholar.org/paper/RoBERTa%3A-A-Robustly-Optimized-BERT-Pretraining-Liu-Ott/077f8329a7b6fa3b7c877a57b81eb6c18b5f87de
189. [R] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer - Reddit, accessed on May 1, 2025, https://www.reddit.com/r/MachineLearning/comments/dm9m33/r_exploring_the_limits_of_transfer_learning_with/
190. arXiv:1910.10683v4 [cs.LG] 19 Sep 2023, accessed on May 1, 2025, https://arxiv.org/pdf/1910.10683
191. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer - Semantic Scholar, accessed on May 1, 2025, https://www.semanticscholar.org/paper/Exploring-the-Limits-of-Transfer-Learning-with-a-Raffel-Shazeer/6c4b76232bb72897685d19b3d264c6ee3005bc2b
192. DL-NLP-Readings/Bibtex/RoBERTa - A Robustly Optimized BERT Pretraining Approach.bib at master - GitHub, accessed on May 1, 2025, https://github.com/IsaacChanghau/DL-NLP-Readings/blob/master/Bibtex/RoBERTa%20-%20A%20Robustly%20Optimized%20BERT%20Pretraining%20Approach.bib

193. Snapshot-Guided Domain Adaptation for ELECTRA - ACL Anthology, accessed on May 1, 2025, https://aclanthology.org/2022.findings-emnlp.163.pdf
194. (PDF) ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators (2020) | Kevin Clark | 1663 Citations - SciSpace, accessed on May 1, 2025, https://scispace.com/papers/electra-pre-training-text-encoders-as-discriminators-rather-31r39dgsbs
195. google-research/electra: ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators - GitHub, accessed on May 1, 2025, https://github.com/google-research/electra
196. Reproducibility Challenge: ELECTRA (Clark et al. 2020) - Wandb, accessed on May 1, 2025, https://wandb.ai/cccwam/rc2020_electra_pretraining/reports/Reproducibility-Challenge-ELECTRA-Clark-et-al-2020---VmlldzozODYzMjk
197. Clark, K., Luong, M.T., Le, Q.V., et al. (2020) ELECTRA Pretraining Text Encoders as Discriminators Rather than Generators. arXiv2003.10555 - References, accessed on May 1, 2025, https://www.scirp.org/reference/referencespapers?referenceid=3023385
198. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators | BibSonomy, accessed on May 1, 2025, https://www.bibsonomy.org/bibtex/2e29e3666539f125f82807fd79f203341/nosebrain
199. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. - DBLP, accessed on May 1, 2025, https://dblp.org/rec/conf/iclr/ClarkLLM20
200. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators | Request PDF - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/340134249_ELECTRA_Pre-training_Text_Encoders_as_Discriminators_Rather_Than_Generators
201. Learning to Sample Replacements for ELECTRA Pre-Training - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/353486242_Learning_to_Sample_Replacements_for_ELECTRA_Pre-Training
202. arXiv:2402.13130v3 [cs.CL] 3 Oct 2024, accessed on May 1, 2025, https://arxiv.org/pdf/2402.13130
203. ELECTRA: PRE-TRAINING TEXT ENCODERS AS DIS- CRIMINATORS RATHER THAN GENERATORS - OpenReview, accessed on May 1, 2025, https://openreview.net/attachment?id=r1xMH1BtvB&name=original_pdf
204. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators - OpenReview, accessed on May 1, 2025, https://openreview.net/pdf?id=r1xMH1BtvB
205. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators - arXiv, accessed on May 1, 2025, https://arxiv.org/abs/2003.10555
206. [D] T5: Exploring Limits of Transfer Learning with Text-to-Text Transformer | Research Paper Walkthrough : r/MachineLearning - Reddit, accessed on May 1,

2025,
https://www.reddit.com/r/MachineLearning/comments/kvv2s2/d_t5_exploring_limits_of_transfer_learning_with/
207. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity – Related Work – Interesting papers - Alastair Reid, accessed on May 1, 2025, https://alastairreid.github.io/RelatedWork/papers/fedus:arxiv:2021/
208. PaLM: Scaling Language Modeling with Pathways - Journal of Machine Learning Research, accessed on May 1, 2025, https://www.jmlr.org/papers/volume24/22-1144/22-1144.pdf
209. RUCAIBox/LLMSurvey: The official GitHub page for the survey paper "A Survey of Large Language Models"., accessed on May 1, 2025, https://github.com/RUCAIBox/LLMSurvey
210. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer - BibBase, accessed on May 1, 2025, https://bibbase.org/network/publication/raffel-shazeer-roberts-lee-narang-matena-zhou-li-etal-exploringthelimitsoftransferlearningwithaunifiedtexttotexttransformer-2019
211. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer - arXiv, accessed on May 1, 2025, https://arxiv.org/abs/1910.10683
212. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer - Journal of Machine Learning Research, accessed on May 1, 2025, https://jmlr.org/papers/volume21/20-074/20-074.pdf
213. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. - DBLP, accessed on May 1, 2025, https://dblp.org/rec/journals/jmlr/RaffelSRLNMZLL20
214. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, accessed on May 1, 2025, https://www.researchgate.net/publication/336767865_Exploring_the_Limits_of_Transfer_Learning_with_a_Unified_Text-to-Text_Transformer
215. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, accessed on May 1, 2025, https://research.google/pubs/exploring-the-limits-of-transfer-learning-with-a-unified-text-to-text-transformer/
216. Paper Summary: Exploring the Limits of Transfer Learning with a Unified Text-to-text Transformer - queirozf.com, accessed on May 1, 2025, https://queirozf.com/entries/paper-summary-exploring-the-limits-of-transfer-learning-with-a-unified-text-to-text-transformer
217. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity - Journal of Machine Learning Research, accessed on May 1, 2025, https://jmlr.org/papers/volume23/21-0998/21-0998.pdf
218. The Switch Transformer | Towards Data Science, accessed on May 1, 2025, https://towardsdatascience.com/the-switch-transformer-59f3854c7050/
219. arXiv:2101.03961v3 [cs.LG] 16 Jun 2022, accessed on May 1, 2025, https://arxiv.org/pdf/2101.03961
220. QMoE: Sub-1-Bit Compression of Trillion-Parameter Models - MLSys

Proceedings, accessed on May 1, 2025, https://proceedings.mlsys.org/paper_files/paper/2024/file/c74b624843218d9b6713fcf299d6d5e4-Paper-Conference.pdf
221. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity, accessed on May 1, 2025, https://jmlr.org/papers/v23/21-0998.html
222. Scaling Instruction-Finetuned Language Models (Flan-PaLM) - Samuel Albanie, accessed on May 1, 2025, https://samuelalbanie.com/digests/2022-10-scaling-instruction-finetuned-language-models/
223. Natural language processing in the era of large language models - PMC, accessed on May 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10820986/
224. 2022 Google Pathways: An Exploration by Dr Alan D. Thompson, accessed on May 1, 2025, https://s10251.pcdn.co/pdf/2022-Alan-D-Thompson-Pathways-Rev-0c.pdf
225. [2402.06196] Large Language Models: A Survey - arXiv, accessed on May 1, 2025, https://arxiv.org/abs/2402.06196
226. PaLM - Wikipedia, accessed on May 1, 2025, https://en.wikipedia.org/wiki/PaLM
227. PaLM: Scaling Language Modeling with Pathways - alphaXiv, accessed on May 1, 2025, https://www.alphaxiv.org/overview/2204.02311
228. Maarten Bosma - Google Scholar, accessed on May 1, 2025, https://scholar.google.com/citations?user=wkeFQPgAAAAJ&hl=en
229. [2204.02311] PaLM: Scaling Language Modeling with Pathways - arXiv, accessed on May 1, 2025, https://arxiv.org/abs/2204.02311
230. PaLM: Scaling Language Modeling with Pathways - BibBase, accessed on May 1, 2025, https://bibbase.org/network/publication/chowdhery-narang-devlin-bosma-mishra-roberts-barham-chung-etal-palmscalinglanguagemodelingwithpathways-2023
231. PaLM: Scaling Language Modeling with Pathways - Journal of Machine Learning Research, accessed on May 1, 2025, https://jmlr.org/papers/volume24/22-1144/22-1144.pdf
232. Aakanksha Chowdhery - Google Scholar, accessed on May 1, 2025, https://scholar.google.com/citations?user=7KDSCpQAAAAJ&hl=en
233. PaLM: Scaling Language Modeling with Pathways - arXiv, accessed on May 1, 2025, https://arxiv.org/pdf/2204.02311
234. PaLM 2 Technical Report - Google AI, accessed on May 1, 2025, https://ai.google/static/documents/palm2techreport.pdf
235. PaLM: Scaling Language Modeling with Pathways - Semantic Scholar, accessed on May 1, 2025, https://www.semanticscholar.org/paper/PaLM%3A-Scaling-Language-Modeling-with-Pathways-Chowdhery-Narang/094ff971d6a8b8ff870946c9b3ce5aa173617bfb
236. (PDF) LLaMA: Open and Efficient Foundation Language Models - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/368842729_LLaMA_Open_and_Efficient

_Foundation_Language_Models
237. (PDF) LLaMA: Open and Efficient Foundation Language Models (2023) | Hugo Touvron | 6015 Citations - SciSpace, accessed on May 1, 2025, https://scispace.com/papers/llama-open-and-efficient-foundation-language-models-1aeusmr9
238. [2302.13971] LLaMA: Open and Efficient Foundation Language Models - arXiv, accessed on May 1, 2025, https://arxiv.org/abs/2302.13971
239. LLaMA: Open and Efficient Foundation Language Models. - DBLP, accessed on May 1, 2025, https://dblp.org/rec/journals/corr/abs-2302-13971
240. Hugo Touvron - Google Scholar, accessed on May 1, 2025, https://scholar.google.com/citations?user=xImarzoAAAAJ&hl=en
241. Paper page - LLaMA: Open and Efficient Foundation Language Models - Hugging Face, accessed on May 1, 2025, https://huggingface.co/papers/2302.13971
242. Learning Fine-Grained Grounded Citations for Attributed Large Language Models - ACL Anthology, accessed on May 1, 2025, https://aclanthology.org/2024.findings-acl.838.pdf
243. Xavier Martinet - Google Scholar, accessed on May 1, 2025, https://scholar.google.com/citations?user=GLfTggQAAAAJ&hl=en
244. [2302.13971] LLaMA: Open and Efficient Foundation Language Models - ar5iv - arXiv, accessed on May 1, 2025, https://ar5iv.labs.arxiv.org/html/2302.13971
245. LLaMA: Open and Efficient Foundation Language Models. - BibSonomy, accessed on May 1, 2025, https://www.bibsonomy.org/bibtex/03a85d2a0612b9704acf6884edbe60aa
246. Noteworthy LLM Research Papers of 2024 - Sebastian Raschka, accessed on May 1, 2025, https://sebastianraschka.com/blog/2025/llm-research-2024.html
247. Understanding LLM Embeddings: A Comprehensive Guide - IrisAgent, accessed on May 1, 2025, https://irisagent.com/blog/understanding-llm-embeddings-a-comprehensive-guide/
248. (PDF) Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/372278221_Large_Language_Models_A_Comprehensive_Survey_of_its_Applications_Challenges_Limitations_and_Future_Prospects
249. A Survey on Self-Evolution of Large Language Models - DEV Community, accessed on May 1, 2025, https://dev.to/aimodels-fyi/a-survey-on-self-evolution-of-large-language-models-4oen
250. 8 Challenges Of Building Your Own Large Language Model - Labellerr, accessed on May 1, 2025, https://www.labellerr.com/blog/challenges-in-development-of-llms/
251. [2412.03220] Survey of different Large Language Model Architectures: Trends, Benchmarks, and Challenges - arXiv, accessed on May 1, 2025,

https://arxiv.org/abs/2412.03220
252. What are LLM Embeddings? - Aisera, accessed on May 1, 2025, https://aisera.com/blog/llm-embeddings/
253. What are LLM Embeddings? - Iguazio, accessed on May 1, 2025, https://www.iguazio.com/glossary/llm-embeddings/
254. Understanding Large Language Models: A Long But Simple Guide - Mediate.com, accessed on May 1, 2025, https://mediate.com/understanding-large-language-models-a-long-but-simple-guide/
255. Bias and Fairness in Large Language Models: A Survey - MIT Press Direct, accessed on May 1, 2025, https://direct.mit.edu/coli/article/50/3/1097/121961/Bias-and-Fairness-in-Large-Language-Models-A
256. Top Research Papers on LLM Models to Read Now - Paperguide, accessed on May 1, 2025, https://paperguide.ai/papers/top/research-papers-llm-models/
257. The Evolution of Multimodal Model Architectures - arXiv, accessed on May 1, 2025, https://arxiv.org/html/2405.17927v1
258. Analyzing the homerun year for LLMs: the top-100 most cited AI papers in 2023, with all medals for open models. - Zeta Alpha, accessed on May 1, 2025, https://www.zeta-alpha.com/post/analyzing-the-homerun-year-for-llms-the-top-100-most-cited-ai-papers-in-2023-with-all-medals-for-o
259. The history, timeline, and future of LLMs - Toloka, accessed on May 1, 2025, https://toloka.ai/blog/history-of-llms/
260. [1802.05365] Deep contextualized word representations - arXiv, accessed on May 1, 2025, https://arxiv.org/abs/1802.05365
261. Deep Contextualized Word Representations | Request PDF - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/325445489_Deep_Contextualized_Word_Representations
262. ELMo Contextual Word Representations - Wolfram Neural Net Repository, accessed on May 1, 2025, https://resources.wolframcloud.com/NeuralNetRepository/resources/ELMo-Contextual-Word-Representations-Trained-on-1B-Word-Benchmark/
263. Deep Contextualized Word Representations - ACL Anthology, accessed on May 1, 2025, https://aclanthology.org/N18-1202/
264. Deep Contextualized Word Representations. - DBLP, accessed on May 1, 2025, https://dblp.org/rec/conf/naacl/PetersNIGCLZ18
265. Deep contextualized word representations - Papers With Code, accessed on May 1, 2025, https://paperswithcode.com/paper/deep-contextualized-word-representations
266. Deep contextualized word representations for detecting sarcasm and irony - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/334116026_Deep_contextualized_word_representations_for_detecting_sarcasm_and_irony
267. arXiv:1909.00512v1 [cs.CL] 2 Sep 2019, accessed on May 1, 2025,

https://arxiv.org/pdf/1909.00512
268. zhouyonglong/Deep-contextualized-word-representations-Tensorflow - GitHub, accessed on May 1, 2025, https://github.com/zhouyonglong/Deep-contextualized-word-representations-Tensorflow
269. Deep Contextualized Word Representations - Semantic Scholar, accessed on May 1, 2025, https://www.semanticscholar.org/paper/Deep-Contextualized-Word-Representations-Peters-Neumann/3febb2bed8865945e7fddc99efd791887bb7e14f
270. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context - ar5iv - arXiv, accessed on May 1, 2025, https://ar5iv.labs.arxiv.org/html/1901.02860
271. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. - SciSpace, accessed on May 1, 2025, https://scispace.com/papers/transformer-xl-attentive-language-models-beyond-a-fixed-17b1kdkcg4
272. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context, accessed on May 1, 2025, https://aclanthology.org/P19-1285/
273. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. - dblp, accessed on May 1, 2025, https://dblp.org/rec/conf/acl/DaiYYCLS19
274. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context | BibSonomy, accessed on May 1, 2025, https://www.bibsonomy.org/bibtex/22fdc4a961b16b4fa36c61feedbfb82db/nosebrain
275. Transformer-XL Review: Beyond Fixed-Length Contexts - Towards Data Science, accessed on May 1, 2025, https://towardsdatascience.com/transformer-xl-review-beyond-fixed-length-contexts-d4fe1d6d3c0e/
276. arXiv:1901.02860v3 [cs.LG] 2 Jun 2019, accessed on May 1, 2025, https://www.cs.cmu.edu/~jgc/publication/Transformer%20XL.pdf
277. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context, accessed on May 1, 2025, https://www.researchgate.net/publication/330276446_Transformer-XL_Attentive_Language_Models_Beyond_a_Fixed-Length_Context
278. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context | Request PDF, accessed on May 1, 2025, https://www.researchgate.net/publication/335781169_Transformer-XL_Attentive_Language_Models_beyond_a_Fixed-Length_Context
279. transformer-xl: attentive language models beyond a fixed-length context - AMiner, accessed on May 1, 2025, https://static.aminer.cn/misc/pdf/weixin/TRANSFORMER-XL.pdf
280. Yang, Z., Dai, Z., Yang, Y., et al. (2019) Xlnet Generalized Autoregressive Pretraining for Language Understanding. Advances in Neural Information Processing Systems, 32. - References - Scientific Research Publishing, accessed on May 1, 2025, https://www.scirp.org/reference/referencespapers?referenceid=3453151

281. Yang, Z., Dai, Z., Yang, Y., et al. (2019) Xlnet Generalized Autoregressive Pretraining for Language Understanding. Proceedings of Annual Conference on Neural Information Processing Systems 2019 (NeurIPS 2019), Vancouver, 8-14 December 2019, 1-11. - References - Scientific Research Publishing, accessed on May 1, 2025, https://www.scirp.org/reference/referencespapers?referenceid=3475196
282. Zhilin Yang - Google Scholar, accessed on May 1, 2025, https://scholar.google.com.hk/citations?user=7qXxyJkAAAAJ&hl=en
283. XLNet: Generalized Autoregressive Pretraining for Language Understanding. - DBLP, accessed on May 1, 2025, https://dblp.org/rec/conf/nips/YangDYCSL19
284. XLNet: Generalized Autoregressive Pretraining for Language Understanding - arXiv, accessed on May 1, 2025, https://arxiv.org/abs/1906.08237
285. Pretrained Generalized Autoregressive Model with Adaptive Probabilistic Label Clusters for Extreme Multi-label Text Classification, accessed on May 1, 2025, http://proceedings.mlr.press/v119/ye20a/ye20a.pdf
286. (PDF) XLNet: Generalized Autoregressive Pretraining for Language Understanding, accessed on May 1, 2025, https://www.researchgate.net/publication/333892322_XLNet_Generalized_Autoregressive_Pretraining_for_Language_Understanding
287. XLNet: Generalized Autoregressive Pretraining for Language Understanding - arXiv, accessed on May 1, 2025, https://arxiv.org/pdf/1906.08237
288. XLNet: Generalized Autoregressive Pretraining for Language Understanding, accessed on May 1, 2025, https://proceedings.neurips.cc/paper/8812-xlnet-generalized-autoregressive-pretraining-for-language-understanding.pdf
289. XLNet: Generalized Autoregressive Pretraining for Language Understanding - Semantic Scholar, accessed on May 1, 2025, https://www.semanticscholar.org/paper/XLNet%3A-Generalized-Autoregressive-Pretraining-for-Yang-Dai/e0c6abdbdecf04ffac65c440da77fb9d66bb474c
290. Intensive Reading Report, accessed on May 1, 2025, https://home.cse.ust.hk/~cktang/csit6000s/Password_Only/lec16-ust-csit6000s-sp25.pdf
291. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity., accessed on May 1, 2025, https://dblp.org/rec/journals/corr/abs-2101-03961
292. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity, accessed on May 1, 2025, https://www.researchgate.net/publication/348403003_Switch_Transformers_Scaling_to_Trillion_Parameter_Models_with_Simple_and_Efficient_Sparsity
293. [R] Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity - Reddit, accessed on May 1, 2025, https://www.reddit.com/r/MachineLearning/comments/kvlk1j/r_switch_transformers_scaling_to_trillion/
294. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity - cs.Princeton, accessed on May 1, 2025,

https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec16.pdf
295. (PDF) A Survey of Large Language Models (2023) | Wayne Xin Zhao | 1214 Citations, accessed on May 1, 2025, https://scispace.com/papers/a-survey-of-large-language-models-f0td5zxu
296. A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery - ACL Anthology, accessed on May 1, 2025, https://aclanthology.org/2024.emnlp-main.498.pdf
297. (PDF) A Survey of Large Language Models (2023) | Wayne Xin Zhao | 483 Citations, accessed on May 1, 2025, https://typeset.io/papers/a-survey-of-large-language-models-f0td5zxu
298. (PDF) A Survey of Large Language Models - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/369740832_A_Survey_of_Large_Language_Models
299. (PDF) Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/372258530_Large_Language_Models_A_Comprehensive_Survey_of_its_Applications_Challenges_Limitations_and_Future_Prospects
300. A Survey of Large Language Models - arXiv, accessed on May 1, 2025, http://arxiv.org/pdf/2303.18223
301. A Survey of Large Language Models, accessed on May 1, 2025, https://bjpcjp.github.io/pdfs/math/2303.18223-LLM-survey-ARXIV.pdf
302. A Survey of Large Language Models - Semantic Scholar, accessed on May 1, 2025, https://www.semanticscholar.org/paper/A-Survey-of-Large-Language-Models-Zhao-Zhou/0b3904d0e229796aff0bda43bb386513353bc992
303. Switch Transformer Explained - Papers With Code, accessed on May 1, 2025, https://paperswithcode.com/method/switch-transformer
304. Shapeshifter: a Parameter-efficient Transformer using Factorized Reshaped Matrices - NeurIPS, accessed on May 1, 2025, https://proceedings.neurips.cc/paper_files/paper/2021/file/09def3ebbc44ff3426b28fcd88c83554-Paper.pdf
305. A critical review of large language models: Sensitivity, bias, and the path toward specialized AI - MIT Press Direct, accessed on May 1, 2025, https://direct.mit.edu/qss/article/5/3/736/120940/A-critical-review-of-large-language-models
306. The debate over understanding in AI's large language models - PMC - PubMed Central, accessed on May 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10068812/
307. Can Neural Language Models (Learn to) Argue? - DebateLab@KIT, accessed on May 1, 2025, https://debatelab.github.io/journal/critical-thinking-language-models.html
308. Artificial Neural Network Language Models Predict Human Brain Responses to

Language Even After a Developmentally Realistic Amount of Training - MIT Press Direct, accessed on May 1, 2025, https://direct.mit.edu/nol/article/5/1/43/119156/Artificial-Neural-Network-Language-Models-Predict
309. The Debate Over “Understanding” in AI's Large Language Models - YouTube, accessed on May 1, 2025, https://www.youtube.com/watch?v=O5SLGAWSXMw
310. Human-Computer Interaction: 6 Ethical Concerns Around LLMs - Data Science Dojo, accessed on May 1, 2025, https://datasciencedojo.com/blog/human-computer-interaction-and-llms/
311. The Working Limitations of Large Language Models - MIT Sloan Management Review, accessed on May 1, 2025, https://sloanreview.mit.edu/article/the-working-limitations-of-large-language-models/
312. What are large language models (LLMs), why have they become controversial?, accessed on May 1, 2025, https://theamericangenius.com/large-language-models/
313. [2307.10169] Challenges and Applications of Large Language Models - arXiv, accessed on May 1, 2025, https://arxiv.org/abs/2307.10169
314. What problems do Large Language Models (LLMs) actually solve very well? [D] - Reddit, accessed on May 1, 2025, https://www.reddit.com/r/MachineLearning/comments/1gjoxpi/what_problems_do_large_language_models_llms/
315. What are the limitations and potential problems associated with deep learning models such as convolutional neural networks and recurrent neural networks? - Quora, accessed on May 1, 2025, https://www.quora.com/What-are-the-limitations-and-potential-problems-associated-with-deep-learning-models-such-as-convolutional-neural-networks-and-recurrent-neural-networks
316. ChatGPT and large language models: what's the risk? - National Cyber Security Centre, accessed on May 1, 2025, https://www.ncsc.gov.uk/blog-post/chatgpt-and-large-language-models-whats-the-risk
317. Large language models: 6 pitfalls to avoid | The Enterprisers Project, accessed on May 1, 2025, https://enterprisersproject.com/article/2023/5/large-language-models-6-pitfalls-avoid
318. LLMs: The Dark Side of Large Language Models Part 1 | HiddenLayer, accessed on May 1, 2025, https://hiddenlayer.com/innovation-hub/the-dark-side-of-large-language-models/
319. Large Language Models Have Pitfalls for National Security - National Defense Magazine, accessed on May 1, 2025, https://www.nationaldefensemagazine.org/articles/2024/1/2/large-language-models-have-pitfalls-for-national-security
320. Large Language Models in Neurology Research and Future Practice - PMC,

accessed on May 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10752640/
321. The Building Blocks of LLMs: Vectors, Tokens and Embeddings - The New Stack, accessed on May 1, 2025, https://thenewstack.io/the-building-blocks-of-llms-vectors-tokens-and-embeddings/
322. What Are Word Embeddings? | IBM, accessed on May 1, 2025, https://www.ibm.com/think/topics/word-embeddings
323. A Closer Look at Embeddings and Their Use in Large Language Models (LLMs), accessed on May 1, 2025, https://kelvin.legal/embeddings_part2/
324. Recurrent neural network based language model - ISCA Archive, accessed on May 1, 2025, https://www.isca-archive.org/interspeech_2010/mikolov10_interspeech.html
325. Recurrent neural network language model adaptation with curriculum learning | Request PDF - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/269724668_Recurrent_neural_network_language_model_adaptation_with_curriculum_learning
326. (PDF) Recurrent neural network based language model - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/221489926_Recurrent_neural_network_based_language_model
327. Word2vec - Wikipedia, accessed on May 1, 2025, https://en.wikipedia.org/wiki/Word2vec
328. Word2Vec Research Paper Explained - Towards Data Science, accessed on May 1, 2025, https://towardsdatascience.com/word2vec-research-paper-explained-205cb7eecc30/
329. Skip-gram Word2Vec Explained - Papers With Code, accessed on May 1, 2025, https://paperswithcode.com/method/skip-gram-word2vec
330. GloVe Explained | Papers With Code, accessed on May 1, 2025, https://paperswithcode.com/method/glove
331. GloVe Research Paper Explained | Towards Data Science, accessed on May 1, 2025, https://towardsdatascience.com/glove-research-paper-explained-4f5b78b68f89/
332. GloVe: Global Vectors for Word Representation, accessed on May 1, 2025, https://nlp.stanford.edu/projects/glove/
333. T5-2.13inch E-paper - LILYGO, accessed on May 1, 2025, https://lilygo.cc/products/t5-2-13inch-e-paper
334. History, Development, and Principles of Large Language Models-An Introductory Survey, accessed on May 1, 2025, https://arxiv.org/abs/2402.06853
335. Generative artificial intelligence: a historical perspective - Oxford Academic, accessed on May 1, 2025, https://academic.oup.com/nsr/article/12/5/nwaf050/8029900
336. Counterfactual Memorization in Neural Language Models - NeurIPS 2025, accessed on May 1, 2025, https://neurips.cc/virtual/2023/poster/72772
337. Brain Mechanisms in Early Language Acquisition - PMC - PubMed Central,

accessed on May 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC2947444/
338. Training Dynamics of Neural Language Models - Naomi Saphra, accessed on May 1, 2025, https://nsaphra.net/uploads/thesis.pdf
339. Word Acquisition in Neural Language Models - ACL Anthology, accessed on May 1, 2025, https://aclanthology.org/2022.tacl-1.1.pdf
340. Information-Restricted Neural Language Models Reveal Different Brain Regions' Sensitivity to Semantics, Syntax, and Context - MIT Press Direct, accessed on May 1, 2025, https://direct.mit.edu/nol/article/4/4/611/117823/Information-Restricted-Neural-Language-Models
341. Exploring the Limits of Language Modeling - arXiv, accessed on May 1, 2025, https://arxiv.org/pdf/1602.02410
342. The Unreasonable Effectiveness of Recurrent Neural Networks : r/MachineLearning - Reddit, accessed on May 1, 2025, https://www.reddit.com/r/MachineLearning/comments/36s673/the_unreasonable_effectiveness_of_recurrent/
343. Advantage of character based language models over word based - Cross Validated, accessed on May 1, 2025, https://stats.stackexchange.com/questions/216000/advantage-of-character-based-language-models-over-word-based
344. A Neural Probabilistic Language Model, accessed on May 1, 2025, http://papers.neurips.cc/paper/1839-a-neural-probabilistic-language-model.pdf
345. A neural probabilistic language model - Yoshua Bengio - SciSpace, accessed on May 1, 2025, https://scispace.com/papers/a-neural-probabilistic-language-model-v51mwoff5f
346. A Neural Probabilistic Language Model, accessed on May 1, 2025, https://www.jmlr.org/papers/v3/bengio03a.html
347. Sequence to Sequence Learning with Neural Networks - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/319770465_Sequence_to_Sequence_Learning_with_Neural_Networks
348. papers/reviews/sequence-to-sequence-learning-with-neural-networks.md at master - GitHub, accessed on May 1, 2025, https://github.com/abhshkdz/papers/blob/master/reviews/sequence-to-sequence-learning-with-neural-networks.md
349. Sequence to Sequence Learning for Event Prediction - ACL Anthology, accessed on May 1, 2025, https://aclanthology.org/I17-2007.pdf
350. Sequence to Sequence Learning with Neural Networks - NIPS papers, accessed on May 1, 2025, http://papers.neurips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf
351. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation | Request PDF - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/308646556_Google's_Neural_Machine

_Translation_System_Bridging_the_Gap_between_Human_and_Machine_Translation
352. Google Neural Machine Translation - Wikipedia, accessed on May 1, 2025, https://en.wikipedia.org/wiki/Google_Neural_Machine_Translation
353. [1609.08144] Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation - arXiv, accessed on May 1, 2025, https://arxiv.org/abs/1609.08144
354. A Neural Network for Machine Translation, at Production Scale - Google Research, accessed on May 1, 2025, https://research.google/blog/a-neural-network-for-machine-translation-at-production-scale/
355. ELMo Explained - Papers With Code, accessed on May 1, 2025, https://paperswithcode.com/method/elmo
356. DEEP CONTEXTUALIZED WORD REPRESENTATIONS - OpenReview, accessed on May 1, 2025, https://openreview.net/pdf?id=S1p31z-Ab
357. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context - arXiv, accessed on May 1, 2025, https://arxiv.org/abs/1901.02860
358. Transformer-XL Explained | Papers With Code, accessed on May 1, 2025, https://paperswithcode.com/method/transformer-xl
359. NeurIPS 2019 XLNet: Generalized Autoregressive Pretraining for Language Understanding Oral, accessed on May 1, 2025, https://neurips.cc/virtual/2019/oral/15838
360. [1702.01923] Comparative Study of CNN and RNN for Natural Language Processing - arXiv, accessed on May 1, 2025, https://arxiv.org/abs/1702.01923
361. A Survey on Self-Evolution of Large Language Models - DEV Community, accessed on May 1, 2025, https://dev.to/mikeyoung44/a-survey-on-self-evolution-of-large-language-models-4oen
362. Tracing the Influence of Large Language Models across the Most Impactful Scientific Works, accessed on May 1, 2025, https://www.mdpi.com/2079-9292/12/24/4957
363. The Evolving Landscape of Large Language Model (LLM) Architectures - re:cinq, accessed on May 1, 2025, https://re-cinq.com/blog/llm-architectures
364. What are the best papers and resources for LLM Development? : r/LocalLLaMA - Reddit, accessed on May 1, 2025, https://www.reddit.com/r/LocalLLaMA/comments/1bghxlq/what_are_the_best_papers_and_resources_for_llm/
365. Foundational must read GPT/LLM papers - Community - OpenAI Developer Forum, accessed on May 1, 2025, https://community.openai.com/t/foundational-must-read-gpt-llm-papers/197003
366. Jordan, M.I. (1986) Serial Order A Parallel Distributed Processing Approach. Institute for Cognitive Science Report 8604, University of California, San Diego. - References - Scientific Research Publishing, accessed on May 1, 2025, https://www.scirp.org/reference/referencespapers?referenceid=1150670
367. serial order: - a parallel distributed - processing approach, accessed on May 1,

2025, https://papers.baulab.info/papers/also/Jordan-1986.pdf
368. Serial order: a parallel distributed processing approach. Technical report, June 1985-March 1986 - OSTI, accessed on May 1, 2025, https://www.osti.gov/biblio/6910294
369. ED276754 - Serial Order: A Parallel Distributed Processing Approach., 1986-May - ERIC, accessed on May 1, 2025, https://eric.ed.gov/?id=ED276754
370. SERIAL ORDER: A PARALLEL DISTRmUTED PROCESSING APPROACH - Computer Science, accessed on May 1, 2025, https://cseweb.ucsd.edu/~gary/PAPER-SUGGESTIONS/Jordan-TR-8604-OCRed.pdf
371. serial order: - a parallel distributed - processing approach - Computer Science, accessed on May 1, 2025, https://cseweb.ucsd.edu/~gary/PAPER-SUGGESTIONS/Jordan-TR-8604.pdf
372. Parallel distributed processing | Request PDF - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/373792998_Parallel_distributed_processing
373. Serial Order: A Parallel Distributed Processing Approach. - DTIC, accessed on May 1, 2025, https://apps.dtic.mil/sti/citations/ADA173989
374. arXiv:2110.15721v2 [cs.CL] 1 Apr 2022, accessed on May 1, 2025, http://arxiv.org/pdf/2110.15721
375. arXiv:1905.03617v3 [cs.NE] 2 Oct 2019, accessed on May 1, 2025, https://arxiv.org/pdf/1905.03617
376. scholar.google.com, accessed on May 1, 2025, https://scholar.google.com/citations?user=Cxi26JcAAAAJ&hl=en
377. Elman, J. L. (1990). Finding Structure in Time. Cognitive Science, 14, 179-211. - References, accessed on May 1, 2025, https://www.scirp.org/reference/referencespapers?referenceid=1092985
378. Jeffrey L. Elman, Finding Structure in Time - PhilPapers, accessed on May 1, 2025, https://philpapers.org/rec/ELMFSI
379. J. Elman, {Finding structure in time} - PhilPapers, accessed on May 1, 2025, https://philpapers.org/rec/ELMFSI-2
380. Finding structure in time - BibSonomy, accessed on May 1, 2025, https://www.bibsonomy.org/bibtex/27399042378ca115d0951674538ded4b8/butz
381. J. L. Elman, "Finding structure in time," Cognitive Science, Vol. 14, No. 2, pp. 179–211, 1990., accessed on May 1, 2025, https://www.scirp.org/reference/referencespapers?referenceid=8120
382. Jeffrey L. Elman - DBLP, accessed on May 1, 2025, https://dblp.org/pid/69/6234
383. Finding structure in time - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/339789954_Finding_structure_in_time
384. Finding Structure in Time - Semantic Scholar, accessed on May 1, 2025, https://www.semanticscholar.org/paper/Finding-Structure-in-Time-Elman/668087f0ae7ce1de6e0bd0965dbb480c08103260
385. Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. Neural Computation, 9, 1735-1780. - References - Scientific Research Publishing,

accessed on May 1, 2025, https://www.scirp.org/reference/referencespapers?referenceid=3685580
386. (PDF) A Neural Probabilistic Language Model - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/221618573_A_Neural_Probabilistic_Language_Model
387. A Neural Probabilistic Language Model - BibSonomy, accessed on May 1, 2025, https://www.bibsonomy.org/bibtex/25bc1d3d1be6247dd6365014919a711ef/dallmann
388. (PDF) A Neural Probabilistic Language Model - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/2413241_A_Neural_Probabilistic_Language_Model
389. A Neural Probabilistic Language Model - NIPS papers, accessed on May 1, 2025, https://papers.nips.cc/paper/1839-a-neural-probabilistic-language-model
390. Yoshua Bengio - Google Scholar, accessed on May 1, 2025, https://scholar.google.com/citations?user=kukA0LcAAAAJ&hl=en
391. A Neural Probabilistic Language Model - CiteSeerX, accessed on May 1, 2025, https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=e92530755e75919ca3ad287ca888a73ff4c22798
392. A Neural Probabilistic Language Model - Journal of Machine Learning Research, accessed on May 1, 2025, https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf
393. A Neural Probabilistic Language Model - Département d'informatique et de recherche opérationnelle, accessed on May 1, 2025, https://www.iro.umontreal.ca/~vincentp/Publications/lm_jmlr.pdf
394. Recurrent neural network based language model. - DBLP, accessed on May 1, 2025, https://dblp.org/rec/conf/interspeech/MikolovKBCK10
395. Tomas Mikolov - Google Scholar, accessed on May 1, 2025, https://scholar.google.com/citations?user=oBu8kMMAAAAJ&hl=en
396. A Parallel Recurrent Neural Network for Language Modeling with POS Tags - ACL Anthology, accessed on May 1, 2025, https://aclanthology.org/Y17-1021.pdf
397. (PDF) Recurrent neural network based language model - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/311469848_Recurrent_neural_network_based_language_model
398. arXiv:1906.03591v2 [cs.CL] 13 Jun 2019, accessed on May 1, 2025, https://arxiv.org/pdf/1906.03591
399. (PDF) Recurrent neural network based language model (2009) | Tomas Mikolov - SciSpace, accessed on May 1, 2025, https://typeset.io/papers/recurrent-neural-network-based-language-model-1hrkhfza3r
400. Recurrent neural network based language model - Semantic Scholar, accessed on May 1, 2025, https://www.semanticscholar.org/paper/Recurrent-neural-network-based-langua

ge-model-Mikolov-Karafi%C3%A1t/9819b600a828a57e1cde047bbe710d3446b30da5
401. CONTEXT DEPENDENT RECURRENT NEURAL NETWORK LANGUAGE MODEL Tomas Mikolov ∗ BRNO University of Technology Czech Republic Geoffre - Microsoft, accessed on May 1, 2025, https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/rnn_ctxt.pdf
402. [1301.3781] Efficient Estimation of Word Representations in Vector Space - arXiv, accessed on May 1, 2025, https://arxiv.org/abs/1301.3781
403. Efficient Estimation of Word Representations in Vector Space - BibBase, accessed on May 1, 2025, https://bibbase.org/network/publication/mikolov-chen-corrado-dean-efficientestimationofwordrepresentationsinvectorspace-2013
404. Efficient estimation of word representations in vector space -Mikolov, Tomas, et al - YouTube, accessed on May 1, 2025, https://www.youtube.com/watch?v=jZgMIukDjXQ
405. Efficient Estimation of Word Representations in Vector Space - arXiv, accessed on May 1, 2025, https://arxiv.org/pdf/1301.3781
406. (PDF) Efficient Estimation of Word Representations in Vector Space - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/234131319_Efficient_Estimation_of_Word_Representations_in_Vector_Space
407. Efficient Estimation of Word Representations in Vector Space - OpenReview, accessed on May 1, 2025, https://openreview.net/forum?id=idpCdOWtqXd60
408. Efficient Estimation of Word Representations in Vector Space - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/319770439_Efficient_Estimation_of_Word_Representations_in_Vector_Space
409. Glove: Global Vectors for Word Representation - ACL Anthology, accessed on May 1, 2025, https://aclanthology.org/D14-1162.pdf
410. Efficient Estimation of Word Representations in Vector Space - Semantic Scholar, accessed on May 1, 2025, https://www.semanticscholar.org/paper/Efficient-Estimation-of-Word-Representations-in-Mikolov-Chen/f6b51c8753a871dc94ff32152c00c01e94f90f09
411. Glove: Global Vectors for Word Representation. - BibSonomy, accessed on May 1, 2025, https://www.bibsonomy.org/bibtex/2a6e77a38c13e374ab250e13ae22993ec/thoni
412. (PDF) Glove: Global Vectors for Word Representation - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/284576917_Glove_Global_Vectors_for_Word_Representation
413. Glove: Global Vectors For Word Representation: January 2014 | PDF | Matrix (Mathematics) - Scribd, accessed on May 1, 2025, https://www.scribd.com/document/505867119/Glove-Global-Vectors-for-Word-Representation

414. GloVe: Global vectors for word representation - Mendeley, accessed on May 1, 2025, https://www.mendeley.com/catalogue/c3393f5a-4bdd-3262-9050-c46e501bbe12/
415. Glove: Global Vectors for Word Representation. - DBLP, accessed on May 1, 2025, https://dblp.org/rec/conf/emnlp/PenningtonSM14
416. GloVe: Global Vectors for Word Representation - ACL Anthology, accessed on May 1, 2025, https://aclanthology.org/D14-1162/
417. GloVe: Global Vectors for Word Representation - Kaggle, accessed on May 1, 2025, https://www.kaggle.com/datasets/rtatman/glove-global-vectors-for-word-representation
418. arXiv:1411.5595v2 [cs.CL] 26 Nov 2014, accessed on May 1, 2025, https://arxiv.org/pdf/1411.5595
419. Word Embedding (III): GloVe (Global Vectors) - Menghan Wang, accessed on May 1, 2025, https://menghan-wang.github.io/posts/2022/10/Word-vector3/
420. Sutskever, I., Vinyals, O. and Le, Q.V. (2014) Sequence to Sequence Learning with Neural Networks. Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, 8-13 December 2014, 3104-3112. - References - Scientific Research Publishing, accessed on May 1, 2025, https://www.scirp.org/reference/referencespapers?referenceid=3808470
421. Sequence to Sequence Learning with Neural Networks - ResearchGate, accessed on May 1, 2025, https://www.researchgate.net/publication/265554383_Sequence_to_Sequence_Learning_with_Neural_Networks
422. [1409.3215] Sequence to Sequence Learning with Neural Networks - arXiv, accessed on May 1, 2025, https://arxiv.org/abs/1409.3215
423. Ilya Sutskever - Google Scholar, accessed on May 1, 2025, https://scholar.google.com/citations?user=x04W_mMAAAAJ&hl=en
424. Sequence to sequence learning with neural networks - BibSonomy, accessed on May 1, 2025, https://www.bibsonomy.org/bibtex/f051dfa019b32d710821fc5b4219bd49
425. Sequence-to-Sequence Learning as Beam-Search Optimization - ACL Anthology, accessed on May 1, 2025, https://aclanthology.org/D16-1137.pdf
426. A systematic review on sequence-to-sequence learning with neural network and its models - International Journal of Electrical and Computer Engineering (IJECE), accessed on May 1, 2025, https://ijece.iaescore.com/index.php/IJECE/article/download/22626/14780
427. Bahdanau, D., Cho, K. and Bengio, Y. (2014) Neural Machine Translation by Jointly Learning to Align and Translate. arXiv 1409.0473. - References - Scientific Research Publishing, accessed on May 1, 2025, https://www.scirp.org/reference/referencespapers?referenceid=3808471
428. Neural Machine Translation by Jointly Learning to Align and Translate - BibBase, accessed on May 1, 2025, https://bibbase.org/network/publication/bahdanau-cho-bengio-neuralmachinetra

nslationbyjointlylearningtoalignandtranslate-2015
429. (PDF) Neural Machine Translation by Jointly Learning to Align and Translate, accessed on May 1, 2025, https://www.researchgate.net/publication/265252627_Neural_Machine_Translation_by_Jointly_Learning_to_Align_and_Translate
430. [1409.0473] Neural Machine Translation by Jointly Learning to Align and Translate - arXiv, accessed on May 1, 2025, https://arxiv.org/abs/1409.0473
431. Neural Machine Translation by Jointly Learning to Align and Translate | BibSonomy, accessed on May 1, 2025, https://www.bibsonomy.org/bibtex/2713375898fd7d2477f6ab6dc3dd66c2c/albinzehe
432. Dzmitry Bahdanau - Google Scholar, accessed on May 1, 2025, https://scholar.google.de/citations?user=Nq0dVMcAAAAJ&hl=en
433. Jointly Learning to Align and Translate with Transformer Models - ACL Anthology, accessed on May 1, 2025, https://aclanthology.org/D19-1453.pdf
434. Neural machine translation by - arXiv, accessed on May 1, 2025, https://arxiv.org/pdf/1409.0473
435. Neural Machine Translation by Jointly Learning to Align and Translate - Semantic Scholar, accessed on May 1, 2025, https://www.semanticscholar.org/paper/Neural-Machine-Translation-by-Jointly-Learning-to-Bahdanau-Cho/fa72afa9b2cbc8f0d7b05d52548906610ffbb9c5
436. Table 2 from Google's Neural Machine Translation System: Bridging the Gap between ... - Semantic Scholar, accessed on May 1, 2025, https://www.semanticscholar.org/paper/Google%27s-Neural-Machine-Translation-System%3A-the-Gap-Wu-Schuster/c6850869aa5e78a107c378d2e8bfa39633158c0c/figure/2
437. Google's Neural Machine Translation System: Bridging the Gap between Human and ... - BibSonomy, accessed on May 1, 2025, https://www.bibsonomy.org/bibtex/16b0f528693f1410385c3449ab2885c53
438. Google's Neural Machine Translation System: Bridging the Gap between Human - DBLP, accessed on May 1, 2025, https://dblp.org/rec/journals/corr/WuSCLNMKCGMKSJL16
439. Bridging the Gap between Human and Machine Translation - BibBase, accessed on May 1, 2025, https://bibbase.org/network/publication/wu-schuster-chen-le-norouzi-macherey-krikun-cao-etal-googlesneuralmachinetranslationsystembridgingthegapbetweenhumanandmachinetranslation-2016
440. Google's Neural Machine Translation System - Semantic Scholar, accessed on May 1, 2025, https://www.semanticscholar.org/paper/Google%27s-Neural-Machine-Translation-System%3A-the-Gap-Wu-Schuster/c6850869aa5e78a107c378d2e8bfa39633158c0c
441. Google's neural machine translation system - arXiv, accessed on May 1, 2025, https://arxiv.org/pdf/1609.08144
442. Google's Neural Machine Translation System: - ProQuest, accessed on May 1,

2025,
https://www.proquest.com/docview/2080906303/?sourcetype=Working%20Papers
443. Solved " Google's neural machine translation system: | Chegg.com, accessed on May 1, 2025,
https://www.chegg.com/homework-help/questions-and-answers/google-s-neural-machine-translation-system-bridging-gap-human-machine-translation-describe-q121274944
444. The Breakthrough of Large Language Models Release for Medical Applications: 1-Year Timeline and Perspectives - PubMed Central, accessed on May 1, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10873461/
445. Ravid Shwartz-Ziv - Google Scholar, accessed on May 1, 2025,
https://scholar.google.co.il/citations?user=SqsLFwMAAAAJ&hl=en
446. Towards Understanding the Role of Attention in Prompt-tuning - Google Research, accessed on May 1, 2025,
https://research.google/pubs/towards-understanding-the-role-of-attention-in-prompt-tuning/
447. History of NLA: Echolalia - Communication Development Center, accessed on May 1, 2025, https://communicationdevelopmentcenter.com/history-research/
448. Understanding Encoder And Decoder LLMs - Ahead of AI - Sebastian Raschka, accessed on May 1, 2025,
https://magazine.sebastianraschka.com/p/understanding-encoder-and-decoder/comments
449. LLM Embeddings Explained: A Visual and Intuitive Guide - a Hugging Face Space by hesamation, accessed on May 1, 2025,
https://huggingface.co/spaces/hesamation/primer-llm-embedding
450. Lecture 12: Word Embeddings and Large Language Models - Introduction to Data Science and Machine Learning, accessed on May 1, 2025,
https://lse-me314.github.io/lecturenotes/ME314_day12.pdf