UNDERSTANDING LARGE LANGUAGE MODELS

Amber L. Solberg[*]

[*] Staff Editor, Georgetown Law Technology Review, Volume 9; Editorial Board, Liberty University Law Review, Volume 18; LL.M. Candidate in Technology Law and Policy, Georgetown University Law Center (2025); J.D., Liberty University School of Law (2024); A.L.M. in Extension Studies, Concentration: Data Science, Harvard University (2021); B.S. Financial Mathematics and Statistics, University of California, Santa Barbara (2018).

I. INTRODUCTION

Since the dawn of human civilization, the evolution of spoken language has laid the foundation for communication upon which all human and technological interactions are constructed. Language furnishes us with the vocabulary, semantic subtleties, and grammatical frameworks essential for conveying complex ideas and concepts. Large Language Models (LLMs) are advanced artificial intelligence (AI) systems that have redefined how machines process and generate human language.[1] LLMs utilize massive datasets, deep neural network architectures, and transformer mechanisms to process and produce text that closely resembles human communication.[2] These capabilities make them highly effective for various natural language processing (NLP) tasks, including text translation, code generation, and chatbots. Since their inception, prominent LLMs like ChatGPT and Google Translate have become indispensable tools for advancing technology that seamlessly integrates with human communication. This Technology Explainer provides a high-level overview of the background, key functionalities, operation, and types of LLMs, as well as a commentary on their limitations and potential.

II. ORIGIN OF LLMS: NATURAL LANGUAGE PROCESSING

Although LLMs have gained notoriety since the launch of ChatGPT in 2021, a subfield of computer science called Natural Language Processing (NLP) has been developing the underlying technology for these algorithms over decades. The emergence of intelligent machines and the pressing demand for language translation after World War II served as a catalyst for the development of NLP as a distinct field of computer science. NLP is a field dedicated to creating systems capable of interacting with and processing human language or language-like data in its written, spoken, or structured forms.[3] Rooted in computational linguistics, NLP shifted the focus from merely understanding the abstract principles underlying human language to engineering systems

---

[1] *What Are Large Language Models (LLMs)?*, ELASTIC, https://www.elastic.co/what-is/large-language-models [perma.cc/N2KR-WMU4] (last visited Dec. 29, 2024).

[2] *Id.*

[3] *See A Complete Guide to Natural Language Processing*, DEEPLEARNING.AI, https://www.deeplearning.ai/resources/natural-language-processing/ [perma.cc/BP5A-P55Z] (last updated Jan. 11, 2023) .

capable of manipulating and processing that language in practical, meaningful ways.[4]

NLP's foundational premise is deceptively simple: enabling computers to process, interpret, and generate human language.[5] However, achieving this goal requires unraveling the complexities of syntax, semantics, and pragmatics, each of which reflects the dynamic, context-dependent qualities of human expression.[6] In its early days, NLP relied heavily on rule-based systems, which encoded linguistic structures as a series of deterministic instructions.[7] While groundbreaking, these systems struggled with the subtleties and variability inherent in human language, often failing when confronted with idioms, context shifts, or grammatical inconsistencies.[8]

The introduction of statistical methods and machine learning marked a transformation in NLP systems—machines could now infer language patterns from vast datasets rather than depending on rigidly defined rules.[9] This approach provided the flexibility necessary for tasks such as sentiment analysis, machine translation, and speech recognition.[10] By leveraging probabilistic models and training algorithms on increasingly larger corpora, NLP systems began to mirror human-like adaptability, albeit within defined constraints.[11]

---

[4] *Id.*

[5] Deepak Bhatt, *Natural Language Processing: Bridging the Gap Between Humans and Machines*, GLOB. TECH. REV. (July 7, 2024), https://www.globaltechnologyreview.com/post/natural-language-processing-bridging-the-gap-between-humans-and-machines [perma.cc/E7VW-M5P7].

[6] Alexander S. Gillis, Ben Lutkevich & Ed Burns, *What is Natural Language Processing (NLP)?*, TECHTARGET (last updated Aug. 2024), https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP [perma.cc/KF22-7YYT].

[7] Seyed Saeid Masoumzadeh, *From Rule-Based Systems to Transformers: A Journey Through the Evolution of Natural Language Processing*, MEDIUM (June 19, 2023), https://medium.com/@masoumzadeh/from-rule-based-systems-to-transformers-a-journey-through-the-evolution-of-natural-language-9131915e06e1# [perma.cc/67WJ-QXEM].

[8] *Id.*

[9] *Id.*

[10] *GenAI vs. LLMs vs. NLP: A Complete Guide*, SCRIBBLEDATA, https://www.scribbledata.io/blog/genai-vs-llms-vs-nlp-a-complete-guide/ [perma.cc/SEX3-SXVP] (last visited Dec. 29, 2024).

[11] *Id.*

III. WHAT IS A LARGE LANGUAGE MODEL (LLM)?

The emergence of LLMs represents the culmination of decades of progress in NLP, introducing a new level of sophistication.[12] LLMs transcend the capabilities of earlier NLP systems by harnessing deep learning—a subset of machine learning designed to mimic the layered processing of information in the human brain.[13] Much like the human brain requires education and refinement, large language models undergo a process of pre-training followed by fine-tuning to adeptly address tasks such as text classification, question answering, document summarization, and text generation.[14] The extensive problem-solving capabilities of LLMs have wide-ranging applications across various fields and provide the basis for a multitude of NLPapplications like machine translation, conversational chatbots, and intelligent virtual assistants.[15]

A. TRANSFORMER MODELS

To enable their sophisticated predictive capabilities, LLMs are supported by an advanced neural network architecture known as the transformer model.[16] Much like the brain, transformers are built upon intricate, multilayered networks comprising countless interconnected nodes, which collectively process information, extract meaning, and generate responses.[17] Transformer models emulate the brain's ability to process external stimuli—they receive inputs, interpret them within a contextual framework, and output coherent responses.[18] This conceptual foundation informs the architecture of transformers, which typically follow an encoder-decoder structure.[19] The encoder's primary role is to deconstruct input sequences into tokens—discrete units of information that can

[12] *Id.*

[13] *What is Natural Language Processing (NLP)?*, AWS, https://aws.amazon.com/what-is/nlp/# [perma.cc/5SPG-LXAX] (last visited Dec. 29, 2024).

[14] *What Are Large Language Models (LLMs)?*, *supra* note 1.

[15] *Id.*

[16] Rick Merritt, *What is a Transformer Model?*, NVIDIA (Mar. 25, 2022), https://blogs.nvidia.com/blog/what-is-a-transformer-model/ [perma.cc/C4UC-ZTX9].

[17] Tyler Au, *An Introduction to the Transformer Model: The Brains Behind Large Language Models*, LYRID.IO (May 1, 2024), https://www.lyrid.io/post/an-introduction-to-the-transformer-model-the-brains-behind-large-language-models# [perma.cc/B2RF-ZYYB].

[18] *Id.*

[19] *What Are Large Language Models (LLMs)?*, *supra* note 1.

represent words, subwords, or characters in the case of NLPs).[20] Through sequential parsing and pattern recognition, the encoder extracts the latent semantic and syntactic relationships embedded within the tokens.[21] The decoder then reconstructs this processed information into an output sequence, effectively transforming raw input into meaningful, contextually relevant language.[22]

B. CORE COMPONENTS OF LLMS

The encoder and decoder themselves consist of multiple transformer layers, each serving a distinct computational function.[23] For instance, attention mechanisms, normalization processes, and feed-forward networks operate in tandem to ensure that the model can both grasp and preserve the nuanced relationships within data.[24] These layers do not work in isolation; rather, they synergistically refine and transmit information through the model's architecture, progressively enhancing the precision of its representations.[25]

1. *Embedding Layer*

The embedding layer serves as the initial interface between raw text and the model's computations. This layer translates discrete textual units, such as words or characters, into dense vector representations known as embeddings.[26] These embeddings encode both the semantic meaning and syntactic relationships of the input, capturing nuances such as word similarity, grammatical roles, and contextual relevance.[27] By embedding linguistic features into a high-

[20] Au, *supra* note 17; Pradeep Menon, *Introduction to Large Language Models and the Transformer Architecture*, MEDIUM (Mar. 9, 2023), https://rpradeepmenon.medium.com/introduction-to-large-language-models-and-the-transformer-architecture-534408ed7e61 [perma.cc/TQ32-QRUB].

[21] Au, *supra* note 17; Menon, *supra* note 20.

[22] Au, *supra* note 17.

[23] Josep Ferrer, *How Transformers Work: A Detailed Exploration of Transformer Architecture*, DATACAMP (Jan. 9, 2024), https://www.datacamp.com/tutorial/how-transformers-work [https://perma.cc/6PW9-3WEN].

[24] *Id.*

[25] *Id.*

[26] *What Are Large Language Models (LLMs)?*, *supra* note 1.

[27] *Id.*

dimensional space, this layer establishes the foundation for the model's understanding of language.[28]

2. *Feedforward Layer*

Building on these embeddings, the feedforward layers (FFNs) execute a series of transformations through fully connected neural networks.[29] These layers are designed to extract increasingly abstract patterns from the input, moving beyond surface-level meanings to uncover higher-order relationships.[30] For example, feedforward layers may identify implicit connections, such as tone or intent, within the text.[31] This capacity to discern abstracted patterns allows the model to contextualize user input and adapt its processing to a wide range of applications, from sentiment analysis to nuanced conversational responses.

3. *Recurrent Layer*

The recurrent layers provide the structural backbone for interpreting sequential data, processing input one token at a time in the order it appears.[32] These layers are particularly adept at capturing temporal dependencies, enabling the model to understand how

---

[28] *What is Embedding Layer: LLMs Explained*, CHATGPT GUIDE (last updated June 12, 2024), https://www.chatgptguide.ai/2024/02/29/what-is-embedding-layer-llms-explained/ [perma.cc/SA6F-ZV2N].

[29] Punyakeerthi BL, *Understanding Feed Forward Networks in Transformers*, MEDIUM (Apr. 29, 2024), https://medium.com/@punya8147_26846/understanding-feed-forward-networks-in-transformers-77f4c1095c67 [perma.cc/9BQU-25KT].

[30] Ian Goodfellow, Yoshua Bengio & Aaron Courville, *Chapter 6: Deep Feedforward Networks in* DEEP LEARNING BOOK 164–223 (2016), https://www.deeplearningbook.org/contents/mlp.html [perma.cc/6TFX-LYPH]; Sandaruwan Herath, *The Feedforward Network (FFN) in The Transformer Model*, MEDIUM (Apr. 19, 2024), https://medium.com/image-processing-with-python/the-feedforward-network-ffn-in-the-transformer-model-6bb6e0ff18db [perma.cc/JSR2-X88M].

[31] *Feedforward Neural Network*, GEERKSFORGEEKRS (last updated June 20, 2024), https://www.geeksforgeeks.org/feedforward-neural-network/ [perma.cc/5RWF-8CAT].

[32] *See What Are Large Language Models (LLMs)?*, *supra* note 1; *see also* Andrej Karpathy, *The Unreasonable Effectiveness of Recurrent Neural Networks*, GITHUB.IO (May 21, 2015), https://karpathy.github.io/2015/05/21/rnn-effectiveness/ [perma.cc/WJQ9-XRJ3].

earlier words influence the meaning of subsequent ones.[33] For example, in complex sentences with subordinate clauses or idiomatic expressions, recurrent layers help preserve the integrity of meaning across the entire input sequence. This sequential processing is crucial for maintaining coherence, particularly in tasks like summarization or text generation.[34]

4. *Attention Mechanism*

The transformative innovation in modern LLMs, however, lies in the attention mechanism. Attention mechanisms allow the model to focus on keywords in a sentence, much like how humans focus on key phrases in conversations.[35] This component evaluates the relationships between all tokens in a sequence, dynamically assigning weights to determine which words or phrases are most critical for the task at hand.[36] For instance, in a passage discussing multiple topics, attention enables the model to isolate the specific context needed to generate an accurate response. This capability ensures that the model produces outputs that are not only contextually rich but also aligned with the user's intent. What makes attention particularly revolutionary is its ability to integrate information across the entire input sequence simultaneously, unlike earlier NLP models such as Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks, which are constrained by their limited capacity to reference distant elements.[37] This global contextual awareness allows transformers to synthesize complex patterns across vast amounts of data, facilitating their ability to generate coherent, contextually appropriate responses.

C. HOW LLMS OPERATE

LLMs are defined by their extraordinary scale, with millions to billions of parameters serving as the internal variables that allow them to predict and generate text.[38] These parameters are what enable

---

[33] *See* Karpathy, *supra* note 32; *see also Large Language Model (LLM)*, GROWTHLOOP (last updated Feb. 28, 2024), https://www.growthloop.com/university/article/llmv [perma.cc/7YGV-RSVA].

[34] *See* Karpathy, *supra* note 32; *see also Large Language Model (LLM)*, *supra* note 33.

[35] *See* Merritt, *supra* note 16.

[36] *See id.*

[37] *See* Menon, *supra* note 20.

[38] Catherine Breslin, *What's a Parameter in an LLM?*, MEDIUM (Jan. 6, 2024), https://catherinebreslin.medium.com/what-is-a-parameter-3d4b7736c81d [perma.cc/QH9K-CHTZ]; Sean Michael Kerner, *What are*

LLMs to process language with a level of fluency and accuracy that was previously unattainable. Building an LLM starts with pre-training, where the model is exposed (in an unsupervised manner) to massive datasets filled with diverse examples of language.[39] During this phase, the model learns the fundamental patterns and structures of language, such as grammar, context, and word relationships.[40] This foundational training equips the model with a broad understanding that can be applied to a variety of tasks, from language translation to content generation. To tailor an LLM for more specific uses, techniques like fine-tuning and prompt tuning are employed.[41] Both approaches tap into the LLM's extensive pre-trained knowledge, making it adaptable to countless applications while preserving the depth of its original training.[42]

1. *Fine-Tuning*

The goal of fine-tuning is to transform a general-purpose model into one that is tailored for a particular application by training it further on a smaller, focused dataset.[43] This approach takes advantage of the extensive knowledge the model has already gained during pre-training, allowing it to handle specialized tasks without starting from scratch, which would be prohibitively expensive and time-consuming for most organizations.

The process begins by introducing the model to a labeled dataset that corresponds to the desired task.[44] For each example in the dataset, the model makes a prediction and compares it to the correct answer (the label).[45] It calculates the difference, or error, between the prediction and the label, which serves as feedback.[46] Using this

---

Large Language Models (LLMs)?, TECHTARGET (last updated May, 2024), https://www.techtarget.com/whatis/definition/large-language-model-LLM [perma.cc/HD5S-QNRA].

[39] *Pre-training in LLM Development*, TOLOKA.AI (Feb. 22, 2024), https://toloka.ai/blog/pre-training-in-llm-development/ [perma.cc/3LJ2-LEMT].

[40] *Id.*

[41] *Prompt Tuning vs. Fine-Tuning—Preferences, Best Practices and Use Cases*, NEXLA, https://nexla.com/ai-infrastructure/prompt-tuning-vs-fine-tuning/ [perma.cc/D88S-LJ9N] (last visited Dec. 29, 2024).

[42] *Id.*

[43] *Id.*

[44] *Fine-tuning Large Language Models (LLMs) in 2024*, SUPERANNOTATE (July 23, 2024), https://www.superannotate.com/blog/llm-fine-tuning [perma.cc/Y694-GPH3].

[45] *Id,*

[46] *Id.*

feedback, the model adjusts its internal parameters, called weights, through an optimization process like gradient descent.[47] Weights that contribute more to the error are adjusted more significantly, while those that have less impact are changed minimally.[48]

This cycle of prediction, error calculation, and weight adjustment repeats over multiple passes through the dataset, known as epochs.[49] With each iteration, the model refines its internal representations, gradually reducing the error and honing its performance on the specific task.[50] By the end of the fine-tuning process, the model has shifted from a broadly trained system to one that is well-suited to the purpose for which it was tuned.

2. *Prompt Tuning*

Prompt tuning is a method for refining the output of LLMs by introducing specialized, adjustable parameters known as soft prompts.[51] Unlike traditional fine-tuning, which modifies the internal parameters of the model, prompt tuning keeps the core architecture and pre-trained weights of the model unchanged.[52]

Soft prompts are artificial tokens represented as trainable vectors that are added to the input sequence before being processed by the model.[53] These prompts act as task-specific cues, guiding the model's responses without altering its internal weights.[54] Soft prompts can be initialized randomly or based on pre-defined heuristics.[55] Once initialized, they are appended to the input data, ensuring that the model interprets both the prompts and the actual input simultaneously.[56]

In the training phase, the combined input—soft prompts plus task-specific data—is passed through the model.[57] During the forward pass, the model processes this input through its layers to generate an output. A loss function is then applied to compare the model's output to the expected results, calculating the discrepancy

---

[47] *Id.*
[48] *Id.*
[49] *Id.*
[50] *Id.*
[51] *Prompt Tuning vs. Fine-Tuning*, *supra* note 41.
[52] *Id.*
[53] Dimitri Didmanidze, *Understanding Prompt Tuning: Enhance Your Language Models with Precision*, DATACAMP (May 19, 2024), https://www.datacamp.com/tutorial/understanding-prompt-tuning [perma.cc/CBS4-C8H5].
[54] *Id.*
[55] *Id.*
[56] *Id.*
[57] *Id.*

or "error."[58] This loss serves as the guiding metric for improving the soft prompts. Backpropagation, a standard optimization method in neural networks, is used to update parameters.[59] However, in prompt tuning, only the parameters of the soft prompts are adjusted; the model's core weights remain untouched.[60] The errors are propagated backward through the network, and the soft prompts are fine-tuned to better align the model's output with the desired outcome.[61]

The cycle of forward passes, loss evaluation, and backpropagation is repeated over multiple epochs.[62] With each iteration, the soft prompts adapt further, learning how to shape the input in a way that minimizes errors and maximizes task-specific performance.[63] Over time, this iterative process allows the model to become highly specialized for the task at hand while preserving its general-purpose functionality.

IV. BENEFITS AND LIMITATIONS OF LLMS

A. BENEFITS AND USE CASES

It is undisputed that LLMs have revolutionized how machines process, interpret, and generate human language with unprecedented sophistication. These models excel at understanding nuanced contexts, recognizing complex linguistic patterns, and generating text that is both contextually relevant and human-like in quality—making them invaluable for businesses and individual users alike. But even beyond their linguistic capabilities, LLMs have become essential tools for streamlining processes and driving innovation across industries such as healthcare, finance, education, entertainment, customer service, and software development.[64]

In customer service and support, LLMs power virtual assistants and chatbots capable of delivering highly personalized and context-aware responses, significantly improving user satisfaction.[65] Social

---

[58] *Id.*
[59] *Id.*
[60] *Id.*
[61] *Id.*
[62] *Id.*
[63] *Id.*
[64] Pooja Choudhary, *Benefits And Limitations Of LLM*, AIThority (June 18, 2024), https://aithority.com/machine-learning/benefits-and-limitations-of-llm/ [perma.cc/3JE8-4PNM].
[65] Shyam Achuthan, *How AI-Powered Language Models are Transforming the Customer Support Landscape*, LINKEDIN (Mar. 19, 2024), https://www.linkedin.com/pulse/how-ai-powered-language-models-transforming-customer-support-shyam-njbsf [perma.cc/Z9V3-6QBJ].

media platforms utilize LLMs to analyze user sentiment, predict trends, and generate tailored content, fostering deeper connections between users and platforms.[66] In e-commerce and retail, LLMs facilitate sophisticated recommendation engines and dynamic pricing strategies, enabling businesses to anticipate consumer preferences and optimize sales in real time.[67] Financial institutions leverage LLMs for fraud detection, risk analysis, and predictive modeling, enabling them to identify irregularities in vast, complex datasets with precision and speed.[68] In marketing and advertising, LLMs craft hyper-targeted campaigns and conduct sentiment analysis, ensuring messaging resonates with specific audiences.[69] Meanwhile, in healthcare, LLMs support critical functions such as analyzing patient data, accelerating drug discovery, and aiding in diagnostic decision-making through AI-driven tools.[70] These diverse applications underscore the versatility of LLMs in automating complex processes, uncovering actionable insights, and enhancing decision-making, making them indispensable assets across both established and emerging domains.[71]

B. CHALLENGES

Despite their transformative capabilities, LLMs face significant challenges that limit their reliability, scalability, and

---

[66] Shanthi Kumar V, *How AI LLMs Transform Social Media Interactions*, LINKEDIN (Jan. 31, 2024), https://www.linkedin.com/pulse/how-ai-llms-transform-social-media-interactions-shanthi-kumar-v--fuoff/ [perma.cc/D4QA-GJK6].

[67] *The Role of Large Language Models in eCommerce & Retail Industry in 2024*, AMPLEWORK SOFTWARE (Sept. 30, 2024), https://www.amplework.com/blog/large-language-models-in-ecommerce-and-retail/ [perma.cc/E9G8-DQYJ].

[68] *LLMs in Banking*, PACIFIC DATA INTEGRATORS (Oct. 1, 2024), https://www.pacificdataintegrators.com/blogs/llms-in-banking-enhance-fraud-detection-risk-assessment [perma.cc/PPK3-WU86].

[69] *How to Use Large Language Models for Marketing*, KIRAN VOLETI (Aug. 17, 2024), https://kiranvoleti.com/how-to-use-large-language-models-llms-for-marketing [perma.cc/2KMV-RPJ2].

[70] Shuroug A. Alowais, Sahar S. Alghamdi, Nada Alsuhebany, Tariq Alqahtani, Abdulrahman I. Alshaya, Sumaya N. Almohareb, Atheer Aldairem, Mohammed Alrashed, Khalid Bin Saleh, Hisham A. Badreldin, Majed S. Al Yami, Shmeylan Al Harbi & Abdulkareem M. Albekairy, *Revolutionizing Healthcare: The Role of Artificial Intelligence in Clinical Practice*, 23 BMD MED. ED., 2023 at 1–15, https://doi.org/10.1186/s12909-023-04698-z.

[71] *A Guide to Large Language Models (LLMs) For Enterprises*, DAVE AI (Aug. 2024), https://www.iamdave.ai/blog/a-guide-to-large-language-modelsllms-for-enterprises/ [perma.cc/389V-DLVR].

ethical deployment. One of the primary concerns lies in their dependence on massive datasets that often include inherent biases, inaccuracies, and outdated information.[72] These biases can manifest in outputs, perpetuating harmful stereotypes or generating skewed results that reflect the imperfections of the training data. The sheer scale of LLMs also presents logistical hurdles: their resource-intensive nature demands extensive computational power and energy, making them environmentally taxing and financially prohibitive for smaller organizations.[73] Furthermore, their "black-box" architecture compounds these issues, as the complexity of their internal workings obscures how decisions are made, leaving users with little insight into the reasoning behind incorrect or unexpected outputs.[74]

LLMs also grapple with maintaining factual accuracy and contextual appropriateness, often "hallucinating" incorrect information or failing to verify factual consistency in their responses.[75] This poses substantial risks in high-stakes domains such as healthcare, law, and finance, where precision is paramount. Customizing LLMs for specific applications introduces further complexities, requiring specialized expertise to fine-tune models effectively while avoiding overfitting or the introduction of new biases.[76] Addressing these multifaceted challenges requires a combination of strategies: employing rigorous data curation to minimize bias, developing energy-efficient architectures to reduce resource demands, and incorporating interpretability frameworks to make model behavior more transparent.

V. CONCLUSION

LLMs epitomize the confluence of linguistic theory and computational innovation, pushing the boundaries of what machines can achieve with language. As LLMs continue to evolve, they highlight both the remarkable progress in artificial intelligence and the enduring complexity of human communication—a complexity that remains the ultimate challenge and inspiration for this field.

---

[72] *5 Biggest Challenges with LLMs and How to Solve Them*, TENEO.AI, https://www.teneo.ai/blog/5-biggest-challenges-with-llms-and-how-to-solve-them [perma.cc/B35S-5XZW] (last visited Dec. 29, 2024).

[73] *Id.*

[74] *Id.*

[75] *Id.*

[76] *Id.*