International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V73I9P103

Volume 73 Issue 9, 19-27, September 2025

Review Article

Everyday AI: Real-World Applications of Transformer-Based Language Models

Vaishnavi Visweswaraiah

Independent Researcher and Data/AI professional, Ohio, USA.

Corresponding Author : vvaishnavi2021@outlook.com

Received: 16 July 2025 Revised: 18 August 2025 Accepted: 05 September 2025 Published: 30 September 2025

Abstract *- Artificial Intelligence (AI) systems called Large Language Models (LLMs), powered by transformer architectures, have become integral to everyday digital interactions. Although we may not always notice them, these models are used directly or indirectly on many platforms. Well-known models such as Gemini, GPT, T5, and Llama are used in search engines, recommendation systems, social media, and conversational assistants. This article reviews three transformer architectures (encoder-decoder, encoder-only, and decoder-only) that enable computers to understand language, perform translation, generate content, and provide personalized suggestions by mapping it to real-world implementations, such as YouTube Music, Google Search, DoorDash, Netflix, Uber, and many others by highlighting the underlying models, their associated transformer architectures, and their functionalities. This article aims to promote AI literacy and improve understanding of how LLMs shape daily digital experiences.*

Keywords *- Artificial Intelligence, Large Language Model, Transformer, Architecture, GPT, Llama, Gemini.*

1. Introduction

Artificial Intelligence (AI) is not confined to research labs alone; it has become a silent companion in our everyday lives. Every time we query a search engine, scroll through a social media feed, ask a voice assistant a question, use conversational chatbots, receive movie recommendations, or screen job applications, AI models are working behind the scenes. Large Language Models (LLMs), which are built on transformer architectures, revolutionized how machines process natural human language and make decisions. The introduction of 'Transformers' in 2017 led to the development of Generative Pre-trained Transformer (GPT), Text-to-Text Transfer Transformer (T5), Bidirectional Encoder Representations from Transformers (BERT), Large Language Model Meta AI (Llama), and many more [1, 2]. These models, trained on large text datasets, can interpret natural language, generate human-like responses, and identify behavioural patterns in users.

As a result, current technology providers have integrated these systems into our everyday platforms to enhance the quality and personalization of the services they offer to users. For example, Google's search is powered by an LLM called BERT, and social media platforms like Facebook and Instagram have Meta AI integrated into them, which is powered by an LLM called Llama [3, 4]. Despite the rapid adoption of AI systems, prior research has focused either on the technical design of transformers [1, 5, 6] or specific real-world applications such as healthcare, learning, or smart environments [7, 8, 9,10, 11], or the benefits of AI applications in daily activities [12]. This leaves a research gap in the limited work consolidating technical foundations on how transformer-based LLMs connect to widely used platforms such as search engines, recommendation systems, and social media. This gap creates a problem: there is a limited understanding of how the underlying technologies of LLMs shape real-world, everyday AI applications. As AI rapidly integrates into our lives, understanding the internals of applications used every day is essential.

Unlike existing studies that focus narrowly on technical details or individual domains, this article addresses these gaps through a comprehensive narrative review of transformer architectures and their three variations (encoder-only, decoder-only, and encoder-decoder), illustrating their applications in everyday life, such as search engines, chat assistants, social networks, content recommendation, and more. For each of these applications, we describe the underlying LLM models, the kinds of functionality they enable, and how users experience them, whether knowingly or unknowingly.

2. Related Work

The Introduction of the "Attention is all you need" paper led to various subsequent models, including BERT, GPT, T5, Llama, and many others. Variations of transformer

architecture enable these models to perform multiple tasks, but not limited to text generation, classification, translation, and text predictions [1]. Many previous studies have addressed topics such as large language models and transformers from various perspectives. In this section, we provide details of studies that discuss the use of AI in everyday applications. Some studies focus on the technical foundations of the transformer, and others on real-world applications in specific domains.

On the technical side, articles such as [13] provide the evolution of the model from traditional natural language processing to transformers. Similarly, [5] provides a comparative analysis of three transformer architectures along with their trade-off in the context of training efficiency, Natural Language Understanding (NLU), and generalization capabilities for zero-shot and few-shot learning. Article [6] contributes to the conceptual understanding of LLM architectures, prompt engineering, LLM performance evaluation methods, and improving LLM accuracy with Retrieval-Augmented Generation (RAG). At the same time, [1] provides a systematic catalog of LLM models, including BERT, T5, GPT, and others, along with their underlying transformer architecture variations. These articles provide technical aspects of LLMs and associated transformer variation but remain largely detached from real-world, everyday applications.

On the application side, many articles highlight how LLMs are applied in our daily lives. For instance, articles such as [8] have explored the use of LLMs with applications like ChatGPT to support informal and everyday learning, [14] proposes an LLM-based framework for personalized and accurate travel dairies. Moral and social decisions made by LLM are evaluated in [15]. Every day, productivity and trust issues related to using LLMs for daily activities are discussed in [16].

In the context of domain-specific platforms, smart home applications are another domain where LLMs are utilized. Article [9] highlights how LLMs can act as recognizers of human behavior by analyzing sensor data from smart homes, supporting applications in smart energy and healthcare monitoring. Besides, article [17] introduces a simulation approach where LLMs generate realistic daily schedules, enabled by agent-based modeling of human activity, and article [18] emphasizes conversational AI agents designed to create and manage home automation routines through interactive and gamified interfaces that are powered by GPT-4.

Healthcare is another domain for LLM applications. An article such as [19] applies LLM to therapy as a daily mental health assistant for self-care. Few articles explore LLM-driven conversational agents, such as ChatGPT, for supporting autistic individuals in everyday life [7] and how they impact medical practices, including medical documentation and diagnostic support systems [10].

Finally, review articles such as [20] and [21] explore LLM application trends and challenges, but they fall short of linking transformer architectures to everyday digital applications. To complement the reviewed literature, a bibliometric analysis was conducted by generating term co-occurrence maps using the existing review articles collected from 2020 to 2025, based on the defined query term (resulting in 900+ articles retrieved). These 900+ review articles revealed two clusters of research: (Figures 2 and 3) one focusing on technical aspects of transformers (e.g., architecture, models, attention mechanism, deep learning, natural language processing), and other (Figures 2 and 3 red cluster) centered on LLMs, ChatGPT, Agent-based systems, highlighting their applications, capabilities and domain-specific insights in areas such as healthcare, education and medicine. Additionally, terms related to everyday or daily aspects of AI emerged as separate, weakly connected nodes, confirming that, although the everyday use of LLM is acknowledged in the existing literature, it remains poorly integrated with either technical or application-focused studies.

This bibliometric evidence highlights a research gap, as existing reviews rarely address the AI models and their architectures employed in applications or platform use in everyday life. To understand the technical aspects of AI better, a consolidated article that addresses the technical foundations of LLMs used in real-world applications is needed. This can help stakeholders, policymakers, AI enthusiasts, the public, and industry practitioners to understand the rapidly evolving landscape of AI-powered technologies. To the best of our knowledge, this is the only review article that discusses the consolidated applications of LLMs that we interact with every day in our lives.

3. Background: Transformers and LLMs

Earlier models like Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) were state-of-the-art models for language tasks, but they had two big problems: (i) They process one word at a time, which made training slow and hard to parallelize. (ii) RNN struggled to remember from previous steps as the sequence grew longer, and LSTM, despite handling longer sequences better than RNN, faced issues with very long-range dependencies [2].

To overcome these limitations, transformer architecture was introduced, making training faster as it can process data in parallel and learn relations between all words within a given input sentence simultaneously. This changed the way computers understand human language. This enabled AI to be faster and better at understanding things like translation and conversation, and proved its effectiveness in machine translation tasks [2].

A Transformer is a type of neural network architecture—a brain-inspired model made of artificial neurons (tiny math functions) arranged in layers that are used in Artificial Intelligence (AI) to help computers understand and generate human language. This system uses a "self-attention" mechanism that helps transformers focus on the most important parts and positions within text inputs. For example, "if the sentence 'The dog did not cross the street because it was too tired' is an input sentence to be translated, in this sentence, what does 'it' refer to? Does it refer to the street or the dog? For humans with knowledge of the English language, it is simple, but for an algorithm, it is not. When the transformer is processing, it works on 'it'; the self-attention mechanism allows it to associate 'it' with 'dog [22]."

In recent years, transformers have become the backbone of Large Language Models (LLMs), which are neural networks pre-trained on vast text datasets. This breakthrough in this field is enabled only through transformers combined with increased computational power and the vast availability of training datasets from various sources. These advances have produced LLMs, often referred to as foundation models, such as GPT-3, BERT, and Llama, which achieve near-human-level performance on various tasks, such as answering questions, summarizing information, writing stories, and many other tasks [1]. These foundational models are general-purpose AI systems that can adapt to many applications with minimal fine-tuning or prompt engineering. In short, the transformer's attention mechanism has enabled the current widespread application of LLM that we see in everyday life.

4. Transformer Architectures

Transformers with a self-attention mechanism, consisting of two segments: an encoder and a decoder composed of a stack of 6 identical layers (Figure 1)

An encoder (left half in Figure 1) is the first part of the transformer; It consists of 6 layers stacked on top of each other. Each layer has two parts: the first is a multi-head self-attention mechanism, which allows for capturing different types of relationships between words within a sentence. Second is a position-wise feed-forward network where each word is fully connected to a neural network after self-attention to refine its meaning. To maintain stability and prevent information loss, each sub-layer is wrapped with a residual connection followed by layer normalization**.** In other words, an encoder's job is to read the inputs (e.g., a sentence in English) and understand them. It takes each word and looks at it in the context of the whole sentence. It turns the sentence into a series of numbers (called embedding) that capture the meaning and relation between words [2,5,6].

A decoder (right half in Figure 1) is the second part of the transformer, composed of six identical layers stacked on top of each other. Each decoder layer has three parts. In addition to multi-head self-attention and a feed-forward network similar to an encoder, it also has masked multi-head attention, which is identical to the encoder self-attention mechanism. However, the words are masked so that the model can look into previous words when generating text, rather than future words. Like the encoder, residual connections and normalization are applied around each sub-layer. Overall, the decoder's job is to take the information from the encoder and use it to generate an output (for example, a translated sentence in French). It looks at what the encoder has learned about the input and then generates words one by one to create the output sentence [2,5,6]. It writes or creates the answer/response based on what is encoded and understood. Over time, through research and experimentation with the building of LLM using encoder and decoder concepts, this transformer model branched into three main architectural models: a) encoder-decoder (the original transformer model), b) encoder-only transformer, and c) decoder-only transformer [5].

4.1. Encoder-Decoder Architecture

This is the original Transformer (sequence-to-sequence) model, which utilizes both the encoder and the decoder. Used for tasks that need to take inputs and produce different output sequences, like translation [2], summarization [23,24], or task mapping text-to-text (e.g., paraphrasing, Grammar Correction). Some examples of LLMs based on this architecture include T5, Alexa Teacher Model (AlexaTM 20B [25]), and Bidirectional and Autoregressive Transformer (BART).

4.2. Encoder-Only Architecture

This transformer consists of only an encoder, designed for understanding language, and is used for tasks that do not require generating text, such as classification, entity recognition, sentiment analysis, or question answering [5, 6]. Some examples of LLMs based on this architecture are BERT, Google's ALBERT, and ELECTRA [1, 5].

4.3. Decoder-Only Architecture

Transformer contains only a decoder part, designed to generate text by predicting words one at a time by referring to previously generated words. This architecture type is used for tasks such as autocomplete, text continuation, and chatting [6]. Some examples of LLMs in this category include Llama, Google's Gemini, and GPT, which are decoder-only LLMs powering the most popular applications, such as ChatGPT [5]. In real-world applications, all three architectures are actively used, and these architectures can handle many of the same tasks. The choice of architecture depends on how deeply the input needs to be understood, how the output is generated, and the specific goals or constraints of the system. Additionally, some platforms can even do a combination of these architectures; here, the choice of architecture depends on application requirements, training complexity (e.g., how difficult or time-consuming it is to train the model), and resource consumption (e.g., how much memory, computing power, or energy the model needs) [5].

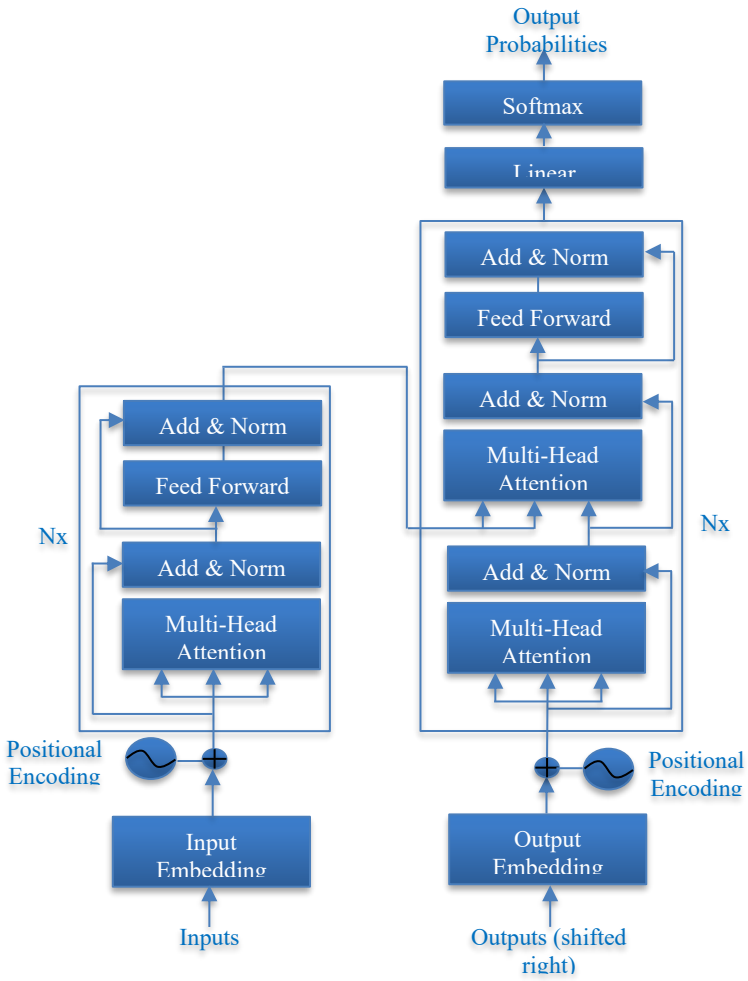

Fig. 1 Transformer model architecture from paper titled "Attention is all you need" [2]

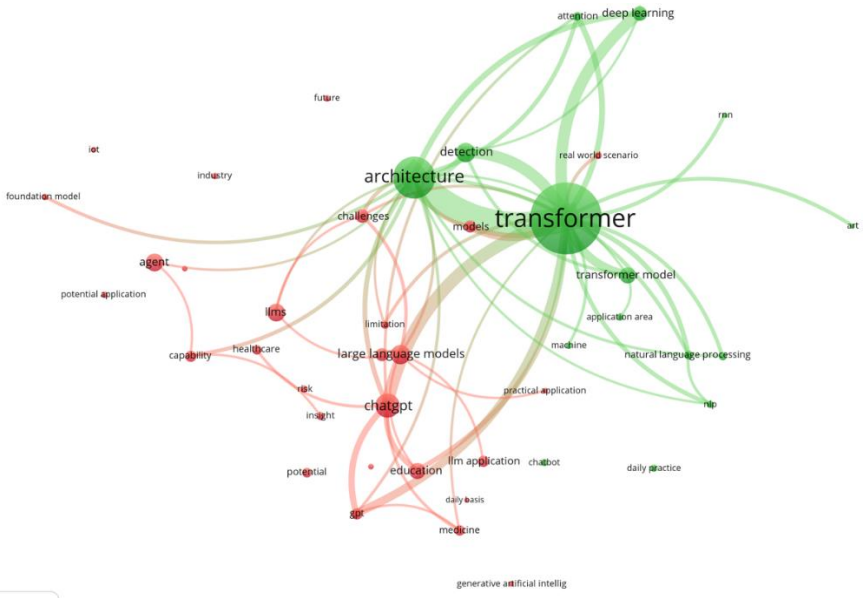

Fig. 2 Network visualization of co-occurring terms from 992 review articles (2020-2025) retrieved using the query ((LLM OR "transformer" OR "transformer architecture") AND (application OR "real-world" OR usage OR platform) AND ("everyday" OR "daily"))

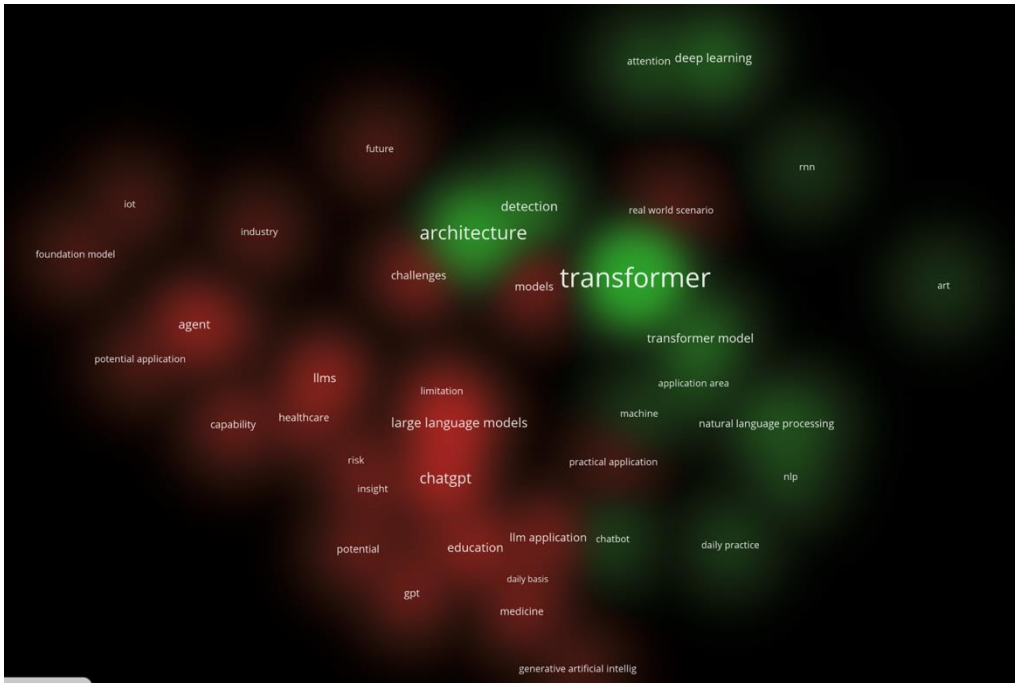

Fig. 3 Density Visualization showing the area of concentrated research focus within the collected articles

5. LLM in Real-World Platforms

The usage of LLMs has changed from research labs to widespread deployment across real-world platforms. This section will provide details about LLMs, the respective architecture pattern used, and the functions they perform in several major application domains, including search engines, recommendation systems, social media, and personal assistants. This section will highlight examples from Google, Netflix, Instagram/Facebook, and YouTube Music. Table 1 provides a high-level overview to complement the detailed platform-specific analysis in this section.

5.1. Google Search

One of the uses of LLM in daily life is Google's search engine. In 2019, Google integrated BERT (an encoder-only transformer model) into its search engine. It understands the full context of sentences by understanding and processing each word in relation to other words within a sentence, enabling better retrieval of information for user search queries [4]. By 2021, Google had introduced the Multitask Unified Model (MUM), which uses T5, an encoder-decoder transformer. MUM can understand and generate language. It also processes and synthesizes information from various source formats, such as text across 75+ languages. This enables Google search to answer complex queries, even if the source information is not in the query's original language [26].

In 2024, Google released Gemini, a multi-modal decoder-only transformer (supporting text, image, audio, video, and advanced coding and more) [27]. Gemini is the foundational LLM for two advanced search features: 1) AI overview, and 2) AI mode.

An "AI overview" is presented at the top of relevant Google search results, providing summaries generated by the Gemini 2.0 model along with sources and citations from which the response is synthesized [28].

"AI mode" helps to dive deep and allows follow-up questions with more advanced reasoning, thinking, and multi-model capabilities. It employs the "query fan-out" technique, which conducts multiple relevant concurrent searches across subtopics and data sources and consolidates the results together [28].

5.2. Facebook and Instagram

In 2019, Meta, the parent company of Facebook and Instagram, leveraged an encoder-only transformer LLM called RoBETa (A Robustly optimized BERT pretraining approach), which is built on Google's BERT to detect hate speech on their social media platforms [29, 30].

Meta also developed its own family of large language models, called Llama, which utilizes the transformer concept for language generation [31]. In 2023, Meta launched Meta AI, powered by Llama 2, an AI assistant that integrates directly into Facebook, Instagram, WhatsApp, and Messenger [32].

By 2024, Meta upgraded the assistant with Llama 3, improving its intelligence and availability, enabling users to access Meta AI to get answers, help plan dinner, generate images, and access real-time information on user feeds, stories, and photos without leaving these applications [3, 33]. These LLMs, Llama 2 and Llama 3, utilize a decoder-only transformer, which serves as the backbone of Meta AI.

5.3. YouTube Music

Recommendation in YouTube Music is powered by both traditional and current state-of-the-art transformers. In general, a recommendation system consists of three main stages: retrieval (collection of relevant items (songs, documents, etc.) from a large corpus), ranking (evaluation of the retrieved results and assigning scores), and filtering (sorting by scores and short-listing) [34]. Google researchers developed a transformer-based ranking system for YouTube Music that captures sequential user actions (plays, skips, likes) and associated contextual information, such as metadata (artists, language of music), time since prior user actions, and the music track associated with the user action. It also incorporates the user's listening history to better predict user preferences [34]. This approach utilizes both existing ranking and transformer-based ranking for YouTube Music recommendations.

5.4. Netflix

A Streaming platform that has begun adopting transformer-based LLMs by shifting from traditional AI algorithms for personalized recommendations. In 2025, Netflix announced a foundation recommendation model for personalized recommendations, replacing many individual Machine Learning (ML) algorithms for distinct needs, including features like "Top Searches", "Trending Now", "Documentaries", "My List", and many other features, with a single LLM-inspired System [35]. This system is specialized for standard behavior rather than natural language tasks, but is similar to GPT-like principles, which is a decode-only style transformer.

5.5. Other Everyday Applications

Beyond Google, YouTube, Netflix, and Meta platforms, transformer-based LLMs have been integrated into numerous daily applications to enhance productivity, communication, and personalized experiences. A few of them are discussed below.

5.5.1. Microsoft 365 (Copilot)

Copilot leverages GPT, a decoder-only LLM to assist users in automating tasks, writing and summarizing content, analyzing and exploring the data, turning ideas into presentations, summarizing emails, drafting a response, researching, and analyzing [36,37,38]. Copilot is integrated into Microsoft 365 apps that are used every day, including Word (for drafting, editing, rewriting, and summarizing documents), Excel (for analyzing data and generating formulas), PowerPoint (for creating and editing presentations), Outlook (for summarizing email threads and drafting emails), Teams (for summarizing meetings and chats, and general follow-ups), and many more [38].

5.5.2. Snapchat (My AI)

Snapchat introduced "My AI" chatbot conversations to answer questions, offer advice on gifts for birthdays, plan a hike, or make dinner suggestions [39]. 'My AI' is powered by decoder-only LLM models like GPT models from OpenAI and Gemini models from Google [40].

5.5.3. LinkedIn

A social networking platform for professional networking and career development, leveraging LLMs such as the InBart model (an in-house domain-adapted encoder-decoder-transformer) for personalized AI-assisted messaging features, GPT-4 (decoder-only transformer) models for LinkedIn Premium profile writing suggestions, and products related to collaborative articles [41].

To enhance LinkedIn's platform further, LinkedIn developed a domain-adapted foundation model called EON, which is a fine-tuned model leveraging Llama 3.18B (decoder-only transformer) [41, 42].

5.5.4. Duolingo

A language learning platform that utilizes LLMs for lesson creation and planning grammar and vocabulary-focused exercises for enrolled customers. The Duolingo Max product features include "Explain my answer" and "Roleplay" functions, powered by GPT-4 (a decoder-only transformer) [43, 44].

5.5.5. Uber

Another everyday app that utilizes LLM for a range of applications, like recommendations and search at Uber Eats, chatbots for customer support services, for code development, and other tasks [45]. Uber powers its applications by leveraging many LLMs available in the market and the AI community, but not limited to open-source models, such as Meta Lama 2 (decoder-only), Mistral AI Mixtral (decoder-only), and proprietary models from Google, OpenAI, and other providers [46, 47].

5.5.6. Grammarly

A writing assistant that integrates LLMs to have advanced intelligent writing and realistic text editing. It uses LLMs to enhance grammar correction, tone/style, and implementation and context-aware suggestions [48]. For collaborative editing, they have fine-tuned CoEdit LLM models using Flan T5, and mEdit (enhanced CoEdit) fine-tuned on LLM models mT5 and mT10 (encoder—decoder) and BLOOMZ, PolyLM, and Bactrian-X (decoder-only models) [49, 50].

5.5.7. Doordash

A food and grocery delivery application leverages both traditional machine learning and Large Language Models (LLMs) for efficient searching features for the best shopping experience for customers and to improve product knowledge graphs, which contain information about products which makes it easier for customers to find exactly what they are looking for, whether it is a specific vintage wine or a flavor of craft beer by leveraging OpenAI's GPT, Google's Bard, and Meta's Llama [51, 52, 53].

6. Conclusion

This article provides a narrative overview of how transformer-based Large Language Models (LLMs) are seamlessly integrated into everyday digital platforms, including Google Search, Facebook, Netflix, Snapchat, and others. By examining the architectures—encoder-only, decoder-only, and encoder-decoder—this paper translates complex AI concepts into a more digestible form for general audiences, aiming to promote AI literacy. We informed the user that foundational models like GPT, Llama, and Gemini are powering search engines, recommendation systems, social media tools, productivity applications, and more. These LLMs, driven by transformers, are no longer confined to academic or enterprise boundaries; they shape the information we consume, the decisions we make, and the experiences we have online.

Table 1. Large Language Models (LLMs) used in everyday digital applications across major technology platforms and architectures employed in them, based on publicly available information

Company	Product / Feature	LLM Model(s) used	Transformer Architecture
Google	Search	BERT, T5 (MUM), Gemini	Encoder-only (BERT), Encoder-Decoder (T5), Decoder-only (Gemini)
Meta	Facebook/Instagram – Hate speech detection, Meta AI	RoBERTa, Llama 2, Llama 3	Encoder-only (RoBERTa), Decoder-only (Llama2/3)
Netflix	Recommendation	Foundation Recommendation Model (GPT-inspired)	Decoder-only (GPT-inspired)
YouTube Music (Google)	Recommendation	Custom transformer-based ranking model	-
Microsoft	Copilot	GPT models	Decoder-only
Snapchat	My AI	GPT, Gemini	Decoder-only
LinkedIn	Messaging, Writing suggestions, Collaborative Articles	InBart, GPT-4, EON (Llama 3.18B)	Encoder-Decoder (InBart), Decoder-only (GPT-4, EON)
Duolingo	Duolingo Max	GPT-4	Decoder-only
Uber	Uber Eats, support chatbots,	Llama 2, Mixtral, models from Google & OpenAI	Decoder-only
Grammarly	Writing assistant	CoEdit (Flan-T5), mEdit (mT5, mT10), BLOOMZ, PolyLM, Bactrian-X	Encoder-decoder (Flan-T5, mT5), Decoder-only (BLOOMZ, PolyLM, Bactrian-X)
DoorDash	Search and product knowledge graph	GPT, Bard, Llama	Decoder-only

Note: Some companies may use additional architectures for other functionalities or tasks. This table reflects only the implementations discussed here.

7. Limitations and Future Work

This work is not a systematic or formal literature review and does not follow standardized protocols for literature selection. The examples discussed in this paper are illustrative rather than exhaustive and are based primarily on publicly available sources such as blog posts and product documentation. Additionally, while the paper focuses on the technical architecture of the model, it does not address the ethical challenges, risks, or biases associated with these systems. With the evolution of AI, it becomes increasingly important for users to understand not only the capabilities of these intelligent systems but also their ethical and societal implications. Future work could explore how such everyday AI systems influence user autonomy, introduce algorithmic bias, and what self-regulatory or policy mechanisms might be needed or in place to ensure responsible deployment.

References

[1] Xavier Amatriain et al., "Transformer Models: An Introduction and Catalog, *arXiv preprint arXiv:2302.07730*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[2] Ashish Vaswani et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Google Scholar]

[3] Meet Your New Assistant: Meta AI, Built with Llama 3, 2024. [Online]. Available: https://about.fb.com/news/2024/04/meta-ai-assistant-built-with-llama-3/

[4] Pandu Nayak, Understanding Searches Better than Ever Before, 2019. [Online]. Available: https://blog.google/products/search/search-language-understanding-bert/

[5] Boyu Liu, "Comparative Analysis of Encoder-Only, Decoder-Only, and Encoder- Decoder Language Models," *International Conference on Data Science and Engineering*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[6] Andrea Filippo Ferraris et al., "The Architecture of Language: Understanding the Mechanics Behind LLMs," *Cambridge forum on AI: Law and governance*, vol. 1, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[7] Dasom Choi et al., "Unlock Life with a Chat(GPT): Integrating Conversational AI with Large Language Models into Everyday Lives of Autistic Individuals," *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Nada Terzimehić, Babette Bühler, and Enkelejda Kasneci, "Conversational AI as a Catalyst for Informal Learning: An Empirical Large-Scale Study on LLM Use in Everyday Learning," arXiv e-prints, arXiv-2506,2025. [CrossRef] [Google Scholar] [Publisher Link]

[9] Gabriele Civitarese et al., "Large Language Models Are Zero-Shot Recognizers for Activities of Daily Living," *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 4, pp. 1–32, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[10] Michael Sonntagbauer, Markus Haar, and Stefan Kluge, "Artificial Intelligence: How will ChatGPT and Other AI Applications Change our Everyday Medical Practice?," *Medical Clinic, Intensive Care and Emergency Medicine*, vol. 118, no. 5, pp. 366–371, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Nina Perry et al., "AI Technology to Support Adaptive Functioning in Neurodevelopmental Conditions in Everyday Environments: A Systematic Review," *npj Digital Medicine*, vol. 7, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Marjona Aslitdinova, "How Artificial Intelligence Helps us in Our Daily Life," *International Journal of Artificial Intelligence*, vol. 1, no. 4, pp. 538–542, 2025. [Google Scholar] [Publisher Link]

[13] Santigo Canchila et al., "Natural Language Processing: An Overview of Models, Transformers and Applied Practices," *Computer Science and Information Systems*, vol. 21, no. 3, pp. 1097–1145, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] Xuchuan Li et al., "Be More Real: Travel Diary Generation Using LLM Agents and Individual Profiles," *arXiv e-prints, arXiv-2407*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] Yu Ying Chiu, Liwei Jiang, and Yejin Choi, "Dailydilemmas: Revealing Value Preferences of LLMs with Quandaries of Daily Life," *arXiv preprint arXiv:2410.02683*, 2024. [CrossRef] [Google Ref] [Publisher Link]

[16] Gaole He, Gianluca Demartini, and Ujwal Gadiraju, "Plan-then-execute: An Empirical Study of user Trust and Team Performance When using LLM Agents as a Daily Assistant," *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1-22, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[17] Haruki Yonekura et al., "Generating Human Daily Activities with LLM for Smart Home Simulator Agents," *2024 International Conference on Intelligent Environments (IE)*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Mathyas Giudici et al., "Designing Home Automation Routines using an LLM-based Chatbot," *Designs*, vol. 8, no. 3, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Jingping Nie et al., "LLM-based Conversational AI Therapist for Daily Functioning Screening and Psychotherapeutic Intervention Via Everyday Smart Devices," *arXiv preprint arXiv:2403.10779*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Wenqi Zheng, and Yutaka Arakawa, "A Review of Wearables-Based Activities of Daily Living Recognition with LLMs: Overview, Progress and Trends," *IEEE Access*, vol. 13, pp. 143813-143830, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[21] Muhammad Usman Hadi et al., "Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects," *Authorea preprints*, vol. 1, no. 3, pp. 1-26. [Google Scholar]

[22] Jay Alammar, The Illustrated Transformer,2018. [Online]. Available: https://jalammar.github.io/illustrated-transformer/

[23] Chintalwar Adhik, Sonti Sri Lakshmi, and C. Muralidharan, "Text Summarization using BART," *AIP Conference Proceedings*, vol. 3075, no. 1. 2024. [CrossRef] [Google Scholar] [Publisher Link]

[24] Glory Odeyemi, "*Finetuning Encoder-Decoder Transformer Models for Text Summarization using Pause Tokens*," University of Windsor, Ontario, Canada, 2025. [Google Scholar] [Publisher Link]

[25] Saleh Soltan et al., AlexaTM 20B: Few-shot Learning using a Large-scale Multilingual seq2seq Model, 2025. [Online]. Available: https://www.amazon.science/publications/alexatm-20b-few-shot-learning-using-a-large-scale-multilingual-seq2seq-model

[26] Pandu Nayak, MUM: A New AI Milestone for Understanding Information, 2021. [Online]. Available: https://blog.google/products/search/introducing-mum/

[27] Sundar Pichai, and Demis Hassabis, Introducing Gemini: Our Largest and Most Capable AI Model, 2023. [Online]. Available: https://blog.google/technology/ai/google-gemini-ai/

[28] Robby Stein, Expanding AI Overviews and Introducing AI Mode, 2025. [Online]. Available: https://blog.google/products/search/ai-mode-search/

[29] AI Advances to Better Detect Hate Speech, 2020. [Online]. Available: https://ai.meta.com/blog/ai-advances-to-better-detect-hate-speech/

[30] Yinhan Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019. [CrossRef] [Google Scholar] [Publisher Link]
[31] Hugo Touvron et al., "LLaMA: Open and Efficient Foundation Language Models," *arXiv* preprint *arXiv:2302.13971*, 2023. [CrossRef] [Google Scholar] [Publisher Link]
[32] Introducing New AI Experiences Across Our Family of Apps and Devices, 2023. [Online]. Available: https://about.fb.com/news/2023/09/introducing-ai-powered-assistants-characters-and-creative-tools/
[33] Introducing Meta Llama 3: The Most Capable Openly Available LLM to Date, 2024. [Online]. Available: https://ai.meta.com/blog/meta-llama-3/
[34] Anushya Subbiah, and Vikram Agarwal, Transformers in Music Recommendation, 2024. [Online]. Available: https://research.google/blog/transformers-in-music-recommendation/
[35] Ko-Jen Hsiao, Yesu Feng, and Sudarshan Lamkhede, Foundation Model for Personalized Recommendation, 2025. [Online]. Available: https://netflixtechblog.com/foundation-model-for-personalized-recommendation-1a0bd8e02d39
[36] Zachary Cavanell, "Microsoft 365 Copilot Wave 2 Spring Updates, 2025. [Online]. Available: https://techcommunity.microsoft.com/blog/microsoftmechanicsblog/microsoft-365-copilot-wave-2-spring-updates/4414785
[37] Microsoft 365 Copilot Release Notes, 2025. [Online]. Available: https://learn.microsoft.com/en-us/copilot/microsoft-365/release-notes
[38] Colette Stallbaumer, Introducing Copilot for Microsoft 365, 2023. [Online]. Available: https://www.microsoft.com/en-us/microsoft-365/blog/2023/03/16/introducing-microsoft-365-copilot-a-whole-new-way-to-work/
[39] Newsroom, Early Insights on My AI. [Online]. Available: https://newsroom.snap.com/early-insights-on-my-ai
[40] What is My AI on Snapchat and How do I use It?. [Online]. Available: https://help.snapchat.com/hc/en-us/articles/13266788358932-What-is-My-AI-on-Snapchat-and-how-do-I-use-it
[41] Praveen Kumar Bodigutla, How we Built Domain-adapted Foundation GenAI Models to Power Our Platform, 2024. [Online]. Available: https://www.linkedin.com/blog/engineering/generative-ai/how-we-built-domain-adapted-foundation-genai-models-to-power-our-platform
[42] Introducing Llama 3.1: Our Most Capable Models to Date, 2024. [Online]. Available: https://ai.meta.com/blog/meta-llama-3-1/
[43] Parker Henry, How Duolingo uses AI to Create Lessons Faster, 2023. [Online]. Available: https://blog.duolingo.com/large-language-model-duolingo-lessons/
[44] Parker Henry, Get to Know the AI Behind Every Video Call with Lily, 2025. [Online]. Available: https://blog.duolingo.com/ai-and-video-call/
[45] OpenAI, Uber Enables Outstanding On-demand Experiences with AI, 2025. [Online]. Available: https://openai.com/index/uber-enables-outstanding-experiences/
[46] Bo Ling et al., Open Source and In-House: How Uber Optimizes LLM Training, 2024. [Online]. Available: https://www.uber.com/en-EG/blog/open-source-and-in-house-how-uber-optimizes-llm-training/
[47] Mistral AI team, Mixtral of Experts, 2023. [Online]. Available: https://mistral.ai/news/mixtral-of-experts
[48] Elevate Your Productivity and Creativity with Generative AI, 2025. [Online]. Available: https://www.grammarly.com/ai/generative-ai
[49] Vipul Raheja et al., "CoEdIT: Text Editing by Task-Specific Instruction Tuning," *arXiv preprint arXiv:2305.09857,* 2023. [CrossRef] [Google Scholar] [Publisher Link]
[50] Vipul Raheja et al., "mEdIT: Multilingual Text Editing via Instruction Tuning," *arXiv preprint arXiv:2402.16472,* 2024. [CrossRef] [Google Scholar] [Publisher Link]
[51] Sudeep Das, Unleashing the Power of Large Language Models at DoorDash for a Seamless Shopping Adventure, 2024. [Online]. Available: https://careersatdoordash.com/blog/unleashing-the-power-of-large-language-models-at-doordash-for-a-seamless-shopping-adventure/
[52] Sissie Hsiao, Bard Becomes Gemini: Try Ultra 1.0 and A New Mobile App Today, 2024. [Online]. Available: https://blog.google/products/gemini/bard-gemini-advanced-app/
[53] Steven Xu, and Sree Chaitanya Vadrevu, Building DoorDash's Product Knowledge Graph with Large Language Models, 2024. [Online]. Available: https://careersatdoordash.com/blog/building-doordashs-product-knowledge-graph-with-large-language-models/