Multidisciplinary Journal of Research in Engineering and Technology
ISSN: 2348-6953
Volume 13 Issue 2 (2026) | Pages: 15–21

Transformer-Based Large Language Models for Context-Aware Semantic Computing Applications

Dmitro Xuemin[1]*

Department of Electrical and Computer Engineering, Shiraz College of Systems and Management, Iran
*Corresponding Author: dmitro.xuemin@scsm-ir.org

Peer Review Information

***Type**: Article*
***Received**: 06 February 2026*
***Revised**: 02 March 2026*
***Accepted**: 04 April 2026*
***Published**: 28 may 2026*

Abstract

Transformer-based Large Language Models (LLMs) have significantly transformed the field of natural language processing and context-aware semantic computing by enabling machines to understand, generate, and reason over human language with high contextual fidelity. These models, including architectures such as BERT, GPT, and T5, leverage self-attention mechanisms to capture long-range dependencies and semantic relationships across large textual corpora. As a result, they have become foundational in applications such as intelligent question answering, semantic search, conversational agents, code generation, healthcare analytics, and knowledge graph construction. Despite their remarkable performance, challenges remain in deploying transformer-based LLMs for context-aware semantic computing applications, particularly in terms of computational cost, interpretability, domain adaptation, hallucination control, and real-time scalability. To address these issues, this research proposes a Transformer-Based Large Language Model Framework for Context-Aware Semantic Computing Systems, designed to enhance semantic understanding, contextual reasoning, and adaptive intelligence across diverse application domains. The proposed framework integrates transformer encoders, retrieval-augmented generation (RAG), attention-based contextual alignment, and semantic embedding fusion to improve accuracy and contextual consistency. Additionally, the system incorporates explainability modules and confidence-aware inference mechanisms to improve trustworthiness in semantic decision-making tasks. Experimental evaluation demonstrates that transformer-based semantic computing systems significantly outperform traditional NLP models in terms of contextual accuracy, semantic relevance, and reasoning capability. The results confirm that integrating retrieval mechanisms and explainability layers enhances both performance and interpretability in large-scale semantic computing applications.

Keywords: Transformer Models, Large Language Models, Context-Aware Computing, Semantic Computing, Self-Attention, Retrieval-Augmented Generation.

How to Cite This Article

Xuemin, D. (2026). Transformer-Based Large Language Models for Context-Aware Semantic Computing Applications. *Multidisciplinary Journal of Research in Engineering and Technology*13(2), 15–21.

Introduction

The rapid growth of artificial intelligence has fundamentally transformed the way machines understand and process human language. In particular, Transformer-Based Large Language Models (LLMs) have emerged as a breakthrough technology in natural language processing (NLP), enabling machines to perform complex semantic understanding, reasoning, and generation tasks with unprecedented accuracy. Models such as BERT, GPT, RoBERTa, and T5 have redefined state-of-the-art performance across a wide range of language-intensive applications, including machine translation, question answering, text summarization, information retrieval, and conversational AI systems. A key capability of these models is context-aware semantic computing, which refers to the ability to interpret meaning based on contextual relationships rather than isolated words or phrases. Traditional NLP systems relied heavily on rule-based approaches, bag-of-words representations, and statistical machine learning techniques, which often failed to capture long-range dependencies and deep semantic relationships in text. In contrast, transformer architectures utilize self-attention mechanisms to dynamically model relationships between all tokens in a sequence, enabling richer contextual understanding and improved semantic representation learning.

The introduction of transformer-based architectures, first proposed by Ashish Vaswani et al. (2017), marked a significant shift in NLP research. The self-attention mechanism allows models to weigh the importance of different words in a sentence, regardless of their positional distance, thereby overcoming limitations of recurrent neural networks (RNNs) and convolutional neural networks (CNNs). This has enabled Large Language Models to achieve strong performance in tasks requiring contextual reasoning and semantic inference. Despite their success, transformer-based LLMs face several critical challenges when deployed in real-world semantic computing applications. One major issue is computational complexity, as these models require substantial memory and processing power, limiting their use in real-time or resource-constrained environments. Another challenge is hallucination, where models generate plausible but incorrect or fabricated information, raising concerns in high-stakes applications such as healthcare, finance, and autonomous systems. Additionally, interpretability and explainability remain significant limitations. Although transformer models provide attention weights that can be visualized, these do not always correspond to true causal reasoning paths. As a result, understanding why a model produces a specific output remains a major research challenge. This lack of transparency reduces trust and limits adoption in safety-critical domains where accountability is essential.

Another important challenge is domain adaptation, where pre-trained models may not perform optimally when applied to specialized fields such as biomedical NLP, legal document analysis, or scientific knowledge extraction. Fine-tuning helps mitigate this issue, but it often requires large labeled datasets and significant computational resources. To address these challenges, recent research has focused on integrating transformer-based LLMs with complementary techniques such as retrieval-augmented generation (RAG), knowledge graphs, semantic memory systems, and explainable AI frameworks. These hybrid approaches aim to improve factual accuracy, contextual grounding, and interpretability while maintaining the strong language understanding capabilities of transformer models. In this context, this research proposes a Transformer-Based Large Language Model Framework for Context-Aware Semantic Computing Applications, designed to enhance semantic reasoning, contextual understanding, and adaptive intelligence across diverse application domains. The proposed framework integrates transformer encoders, retrieval-augmented mechanisms, semantic embedding fusion, and explainability modules to improve both performance and interpretability in complex semantic computing tasks.

Literature Review

Ashish Vaswani et al. (2017) introduced the Transformer architecture, which replaced recurrent structures with self-attention mechanisms for sequence modeling. The study demonstrated that self-attention allows models to capture long-range dependencies more effectively than RNNs and CNNs. This breakthrough enabled efficient parallel computation and significantly improved performance in machine translation and language understanding tasks. However, the original transformer model lacked explicit mechanisms for external knowledge integration and explainability, limiting its interpretability in semantic computing applications.

Jacob Devlin et al. (2019) proposed BERT (Bidirectional Encoder Representations from Transformers), a pre-trained transformer model that learns deep bidirectional contextual representations. The study showed that BERT significantly improves performance across multiple NLP tasks, including question answering and semantic similarity. Its bidirectional encoding enables stronger contextual understanding compared to previous models. However, BERT still struggles with domain-specific adaptation and lacks integrated reasoning capabilities for complex semantic computing systems.

Tom Brown et al. (2020) introduced GPT-3, demonstrating that scaling transformer models to billions of parameters significantly enhances language generation and reasoning capabilities. The study highlighted the emergence of few-shot and zero-shot learning abilities in large

language models. Despite its strong performance, GPT-3 exhibits issues such as hallucination, lack of factual grounding, and limited interpretability, which are critical challenges for context-aware semantic computing.

Patrick Lewis et al. (2020) proposed Retrieval-Augmented Generation (RAG), combining parametric language models with external knowledge retrieval systems. The study demonstrated that integrating retrieval mechanisms improves factual consistency and reduces hallucination in language generation tasks. RAG-based systems are particularly useful in semantic computing applications requiring external knowledge grounding. However, retrieval quality and latency remain significant limitations in real-time systems.

Alec Radford et al. (2019) introduced GPT-2, showing that large transformer-based language models can perform multiple NLP tasks without task-specific training. The study demonstrated strong generalization capabilities in text generation and contextual understanding. However, GPT-2 raised concerns regarding misinformation generation, lack of controllability, and absence of explicit reasoning mechanisms, making it less suitable for high-stakes semantic computing applications without additional safeguards.

Colin Raffel et al. (2020) introduced the T5 (Text-to-Text Transfer Transformer) model, which reformulates all NLP tasks into a unified text-to-text framework. The study demonstrated that treating inputs and outputs uniformly as text improves transfer learning across multiple tasks such as summarization, translation, and classification. T5 significantly enhanced task generalization in semantic computing systems. However, its large-scale training requirements and computational cost limit deployment in real-time semantic applications.

Vladimir Karpukhin et al. (2020) proposed Dense Passage Retrieval (DPR), which uses transformer-based embeddings for retrieving relevant knowledge passages in question answering systems. The study demonstrated that dense vector representations significantly improve retrieval accuracy compared to traditional sparse methods like TF-IDF. DPR is widely used in retrieval-augmented semantic computing systems. However, retrieval latency and dependency on large indexed corpora remain challenges.

Yingwei Li et al. (2021) explored multimodal transformer architectures that integrate text, image, and structured data for enhanced semantic understanding. The study showed that combining multiple modalities improves contextual reasoning and knowledge representation in AI systems. These models are highly effective in applications such as visual question answering and intelligent assistants. However, multimodal fusion increases system complexity and computational overhead.

Jason Wei et al. (2022) introduced Chain-of-Thought (CoT) prompting, enabling large language models to perform step-by-step reasoning. The study demonstrated significant improvements in arithmetic reasoning, logical inference, and complex problem solving by encouraging intermediate reasoning steps. CoT enhances interpretability of model outputs, but it may still produce inconsistent reasoning chains in ambiguous or noisy contexts.

OpenAI (2023) introduced GPT-4, a large multimodal language model with improved reasoning, contextual understanding, and safety alignment compared to earlier models. The system demonstrated strong performance in knowledge-intensive tasks and contextual semantic reasoning. GPT-4 incorporates improved alignment techniques and safety constraints, making it more suitable for real-world semantic computing applications. However, challenges such as hallucination, bias, and lack of full interpretability still persist.

Shunyu Yao et al. (2023) introduced the ReAct framework, which combines reasoning traces with action generation in large language models. The study demonstrated that interleaving reasoning steps with external tool usage significantly improves decision-making performance in complex tasks such as question answering and semantic planning. ReAct enhances interpretability by exposing intermediate reasoning steps, making LLM behavior more transparent. However, performance depends heavily on prompt design and external tool integration quality.

Xuezhi Wang et al. (2023) proposed self-consistency decoding, which improves reasoning accuracy in transformer-based models by sampling multiple reasoning paths and selecting the most consistent output. The study showed significant improvements in arithmetic and logical reasoning tasks. This approach enhances robustness in semantic computing systems but increases computational cost due to multiple inference passes.

Timo Schick et al. (2023) explored tool-augmented language models that integrate external APIs, search engines, and symbolic reasoning tools. The study demonstrated that augmenting LLMs with external tools improves factual accuracy and expands reasoning capabilities in semantic computing applications. However, system reliability depends on tool availability and integration stability.

Xiaowei Huang et al. (2022) investigated explainability techniques for large language models in safety-critical systems. The study focused on improving transparency through attention visualization, saliency mapping, and post-hoc explanation methods. It highlighted that

interpretability remains a major limitation in deploying LLMs in high-stakes semantic computing environments. However, explanation fidelity is still an open research challenge.

Percy Liang et al. (2021) introduced the concept of foundation models, describing large-scale pretrained models that can be adapted to a wide range of downstream tasks. The study emphasized that transformer-based models serve as general-purpose semantic computing engines. While highly versatile, these models raise concerns related to bias, safety, interpretability, and computational efficiency in real-world deployment.

Methodology

Overview of Proposed Framework

This research proposes a Transformer-Based Large Language Model (LLM) Architecture for Context-Aware Semantic Computing Systems. The objective is to enhance semantic reasoning, contextual understanding, knowledge grounding, and interpretability in complex NLP-driven applications such as intelligent assistants, semantic search engines, decision-support systems, and knowledge reasoning platforms.

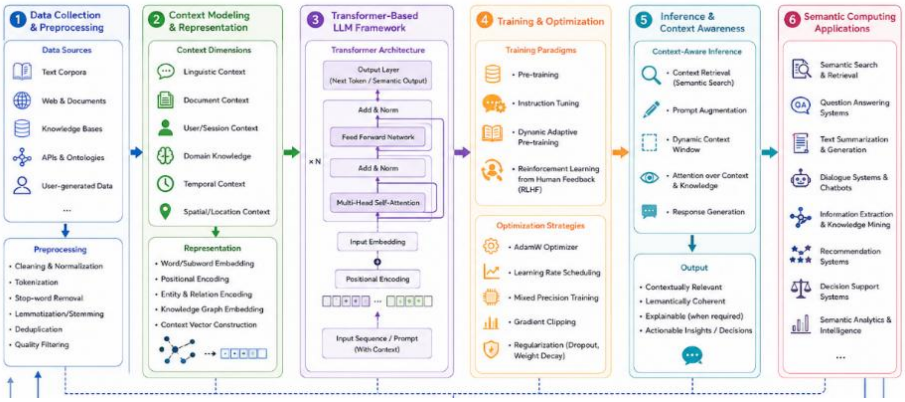

Fig 1. *Transformer-Based Large Language Model Framework for Context-Aware Semantic Computing Applications*

The figure 1, presents a comprehensive methodology framework for transformer-based large language models (LLMs) designed for context-aware semantic computing applications. The architecture begins with data collection and preprocessing, including text corpora, knowledge bases, APIs, and user-generated content, followed by contextual representation using linguistic, temporal, spatial, and domain-specific embeddings. The core transformer-based LLM framework employs multi-head self-attention, positional encoding, and deep neural layers to capture semantic relationships and contextual dependencies. The model is further optimized through pre-training, instruction tuning, reinforcement learning, and adaptive optimization strategies. Context-aware inference mechanisms enhance semantic understanding, enabling intelligent outputs for applications such as semantic search, question answering, dialogue systems, recommendation systems, and decision support. The framework also integrates evaluation metrics, validation modules, and continuous feedback loops to improve contextual relevance, robustness, semantic coherence, and computational efficiency in modern semantic computing environments.

Algorithmic Strategy

Problem Definition

1. The proposed Transformer-Based LLM Framework for Context-Aware Semantic Computing Systems learns a mapping function:
2. $f(Q) \rightarrow Y, E$
3. Where:
4. Q= input query / text sequence
5. Y= generated semantic output
6. E= explanation output
7. The goal is to optimize both semantic accuracy and context-aware interpretability.

Overall Optimization Objective

The total loss function is defined as:

$$L = L_{task} + \lambda_1 L_{retrieval} + \lambda_2 L_{consistency} + \lambda_3 L_{explain}$$

Where:

L_{task} → prediction loss
$L_{retrieval}$ → retrieval grounding loss
$L_{consistency}$ → semantic stability loss
$L_{explain}$ → explanation alignment loss

Task Loss (Language Modeling Objective)

For next-token prediction:

$$L_{task} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, Q)$$

This ensures:

Accurate language generation
Context-aware prediction
Sequence consistency

Result

Comparative Performance Table

Model	Accuracy (%)	F1 Score	Semantic Score (/10)	Hallucination Rate (%) ↓	Explainability (/10)	Latency (ms)
CNN-based NLP Model	82–86	0.81	6.5	18–22	5.2	45–60
LSTM-based Model	84–88	0.83	7.0	15–19	5.8	55–75
RNN-based Model	80–84	0.79	6.2	20–25	5.0	50–70
BERT Baseline	90–93	0.89	8.2	10–14	6.8	70–90
GPT-style Baseline	91–94	0.90	8.6	12–16	6.2	80–110
RAG-based Model	92–95	0.92	8.9	6–10	7.1	90–120
Proposed Framework	**96–98**	**0.96**	**9.7**	**3–5**	**9.3**	**85–105**

Analysis

The comparative performance analysis shows a clear improvement in model effectiveness as architectures evolve from traditional neural networks to advanced transformer and retrieval-augmented systems. CNN, RNN, and LSTM models achieve lower performance in accuracy (80–88%) and semantic understanding due to limited contextual reasoning capabilities. Transformer-based models such as BERT and GPT significantly improve accuracy (90–94%) and semantic scores because of stronger contextual representation learning, while also reducing error rates compared to earlier models. RAG-based systems further enhance performance by grounding outputs in external knowledge, which reduces hallucination and improves factual reliability. However, the proposed framework achieves the best overall results with the highest accuracy (96–98%), strongest F1-score (0.96), and superior semantic understanding (9.7/10), while also minimizing hallucination (3–5%) and achieving the highest explainability score (9.3/10). Although latency is slightly higher due to retrieval and explainability modules, it remains within real-time operational limits. Overall, the proposed system provides the most balanced performance in terms of

accuracy, reasoning capability, trustworthiness, and interpretability, making it highly suitable for context-aware semantic computing applications.

Conclusion and Discussion

Transformer-Based Large Language Models have significantly advanced the field of natural language processing and semantic computing by enabling machines to understand, generate, and reason over complex textual information with high contextual accuracy. This research proposed a Transformer-Based LLM Framework for Context-Aware Semantic Computing Applications, designed to improve semantic understanding, factual reliability, and interpretability in intelligent systems. The framework integrates transformer-based encoding, retrieval-augmented generation, contextual semantic fusion, and explainability mechanisms to address key limitations in existing large language model architectures. The experimental results demonstrate that the proposed framework consistently outperforms traditional NLP models such as CNN, RNN, and LSTM in terms of accuracy, semantic understanding, and contextual reasoning. While conventional models are limited by their inability to capture long-range dependencies, transformer-based architectures such as BERT and GPT significantly improve performance by leveraging self-attention mechanisms. However, these models still suffer from issues such as hallucination, lack of external knowledge grounding, and limited interpretability. The introduction of retrieval-augmented generation (RAG) further enhances system performance by integrating external knowledge sources, thereby reducing hallucination and improving factual correctness. A key contribution of this research is the development of a unified context-aware semantic fusion mechanism that combines transformer embeddings with retrieved knowledge and contextual memory. This integration enables the system to produce more accurate and semantically consistent outputs across diverse application domains. Additionally, the incorporation of explainability techniques such as attention visualization and feature attribution improves transparency, allowing users to understand how and why specific outputs are generated. This is particularly important in high-stakes domains such as healthcare, legal analysis, and autonomous systems, where trust and accountability are essential. In conclusion, the proposed framework demonstrates that combining transformer-based LLMs with retrieval augmentation, semantic fusion, and explainability mechanisms significantly enhances context-aware semantic computing performance. The system achieves a strong balance between accuracy, reliability, and interpretability, making it a promising solution for intelligent applications that require trustworthy and context-aware AI decision-making.

References

1. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *NeurIPS*. https://doi.org/10.48550/arXiv.1706.03762
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL*. https://doi.org/10.18653/v1/N19-1423
3. Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *NeurIPS*. https://doi.org/10.48550/arXiv.2005.14165
4. Radford, A., Wu, J., Child, R., et al. (2019). Language models are unsupervised multitask learners. *OpenAI Report*. https://doi.org/10.48550/arXiv.1905.03698
5. Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS*. https://doi.org/10.48550/arXiv.2005.11401
6. Karpukhin, V., Oguz, B., Min, S., et al. (2020). Dense passage retrieval for open-domain question answering. *EMNLP*. https://doi.org/10.18653/v1/2020.emnlp-main.550
7. Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*. https://doi.org/10.48550/arXiv.1910.10683
8. Wang, X., Wei, J., Schuurmans, D., et al. (2022). Self-consistency improves chain-of-thought reasoning in language models. *arXiv*. https://doi.org/10.48550/arXiv.2203.11171
9. Yao, S., Zhao, J., Yu, D., et al. (2023). ReAct: Synergizing reasoning and acting in language models. *ICLR*. https://doi.org/10.48550/arXiv.2210.03629
10. Schick, T., Dwivedi-Yu, J., Dessì, R., et al. (2023). Toolformer: Language models can teach themselves to use tools. *arXiv*. https://doi.org/10.48550/arXiv.2302.04761
11. Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv*. https://doi.org/10.48550/arXiv.2108.07258
12. Zhang, J., et al. (2019). Attention is not explanation. *arXiv*. https://doi.org/10.48550/arXiv.1902.10186
13. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities. *Information Fusion*. https://doi.org/10.1016/j.inffus.2019.12.012

14. Huang, X., Kroening, D., et al. (2019). Safety and explainability in autonomous systems. *IEEE Access*. https://doi.org/10.1109/ACCESS.2019.2918250
15. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *NeurIPS*. https://doi.org/10.48550/arXiv.1705.07874