IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

Rethinking GPT Intelligence: Towards Evidence-Grounded and Trustworthy Generative AI

Tikone Vaishnavi Ganesh[1], Prof. V. N. Dhakane[2], Prof. P. P. Randive[3], Dr. H. B. Jadhav[4]
[1]Student, Department of Computer Engineering
[2,3]Assistant Professor, Department of Computer Engineering
[4]HOD, Department of Computer Engineering
Adsul's Technical Campus, Chas, Ahilyanagar, India

Abstract: *Generative Pre-trained Transformers (GPTs) have transformed artificial intelligence by enabling systems to generate human-like text, answer questions, summarize information, assist with programming, and support decision-making. Despite these capabilities, the reliability of GPT-based systems remains a significant concern. Fluent responses may contain factual errors, unsupported claims, biased information, or misleading citations. These limitations become particularly important when generative AI is applied to education, healthcare, business, scientific research, and public services. This paper critically reviews selected research published between 2024 and 2026 to examine the technological foundations, practical applications, reliability challenges, and emerging directions of GPT-based systems. The review considers transformer architectures, pre-training, post-training, retrieval-augmented generation, context engineering, model evaluation, and responsible deployment. It identifies a need to evaluate GPT applications through an integrated approach that considers factual accuracy, evidence quality, uncertainty communication, security, computational efficiency, and human oversight. To address this need, the paper proposes a conceptual Human-Centred Evidence-Grounded GPT Framework (HEG-GPT). The framework combines context analysis, trusted-source retrieval, response generation, claim verification, risk assessment, and human review. Unlike an experimentally validated system, this framework represents a proposed research direction whose effectiveness must be established through controlled evaluation. The paper concludes that the future of GPT should focus not only on generating coherent responses but also on producing verifiable, appropriately qualified, secure, and useful outputs.*

Keywords: Generative Pre-trained Transformer, GPT, trustworthy artificial intelligence, large language models, evidence-grounded generation, retrieval-augmented generation, hallucination detection, human-AI collaboration, responsible AI.

I. INTRODUCTION

Artificial intelligence has developed rapidly from systems designed for narrowly defined tasks to generative models capable of producing text, code, images, and other forms of content. Among these developments, Generative Pre-trained Transformers have attracted considerable attention because they can support a wide range of language-related activities through a common model architecture.

The transformer architecture introduced by Vaswani et al. [1] established an important foundation for modern language models. GPT-style models generally use decoder-only transformer architectures and are trained to predict subsequent tokens from preceding context. Large-scale pre-training enables them to learn linguistic patterns and develop capabilities that can be adapted to different tasks.

These capabilities have created opportunities in education, healthcare, software engineering, business communication, research, and customer service. However, practical usefulness depends on more than the fluency of generated text. A

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

response may sound convincing while containing false information, relying on outdated knowledge, or presenting unsupported conclusions as established facts.

Research published in recent years has increasingly examined these limitations. Yenduri et al. [2] provided a broad review of GPT technologies, applications, challenges, and future directions in 2024. Research in 2025 explored model evaluation, efficiency, retrieval-augmented generation, and the responsible adoption of generative AI [3–5]. Studies published in 2026 have continued to examine evaluation across application domains, language-model development, and evidence-grounded systems [6–8].

These developments suggest that GPT should be evaluated as part of a complete application rather than as an isolated language model. Information retrieval, source quality, security controls, user expectations, and human oversight can all influence the reliability of the final result.

This paper therefore focuses on the relationship between GPT capabilities and trustworthy deployment. It reviews selected recent studies, identifies limitations that remain relevant to practical applications, and proposes a conceptual framework for evidence-grounded and human-centred GPT systems.

1. Research objectives

The main objectives of this paper are:

1. To explain the technologies that enable GPT-based systems.
2. To examine the opportunities and limitations of GPT across different application domains.
3. To compare selected research published in 2024, 2025, and 2026.
4. To identify research gaps related to factual reliability, evidence verification, uncertainty, security, and human oversight.
5. To propose a framework connecting evidence retrieval, response generation, verification, and risk-sensitive review.
6. To recommend evaluation methods for future trustworthy GPT applications.

II. RESEARCH METHODOLOGY

This study adopts a focused comparative narrative-review methodology. It examines selected scholarly publications concerning GPT, large language models, generative AI evaluation, retrieval-augmented generation, and responsible AI deployment.

The literature is organized into three publication periods: 2024, 2025, and 2026. The selected studies are examined according to their research objectives, technical focus, practical implications, and relevance to the proposed framework.

The review draws on identifiable journal publications and established foundational research. Priority is given to studies that discuss GPT technologies, model evaluation, reliability, retrieval, or deployment challenges.

1. Analytical approach

The selected studies are compared across five dimensions:

- **Technical foundations:** Model architecture, training, adaptation, and information retrieval.
- **Applications:** The practical tasks and domains supported by generative AI.
- **Reliability:** Factual accuracy, evidence quality, and uncertainty handling.
- **Operational considerations:** Computational efficiency, privacy, security, and monitoring.
- **Human factors:** Oversight, accountability, and the effects of AI assistance on users.

The comparison is qualitative. It does not combine accuracy scores from different papers because the studies may use different datasets, models, tasks, and evaluation procedures.

2. Scope and limitations

This paper is a focused narrative review rather than an exhaustive systematic literature review. It does not claim to have screened every publication from 2024–2026 or followed a complete PRISMA-based selection process. Consequently,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

its findings should be interpreted as a critical synthesis of selected research rather than a statistically representative account of the entire field.

A future systematic study should document database search strings, screening criteria, excluded publications, quality assessments, and the number of papers selected at each stage.

III. TECHNOLOGICAL FOUNDATIONS OF GPT

1. Transformer architecture and attention

The transformer architecture uses attention mechanisms to model relationships between tokens in a sequence. In GPT-style systems, causal attention allows the model to use preceding context when predicting the next token.

The model generates a response progressively, selecting subsequent tokens based on its learned parameters, the prompt, and any additional context supplied during inference.

This approach supports flexible text generation and task adaptation. Nevertheless, attention mechanisms do not independently establish whether a statement is true. Factual reliability also depends on the training data, available evidence, model behavior, and verification procedures.

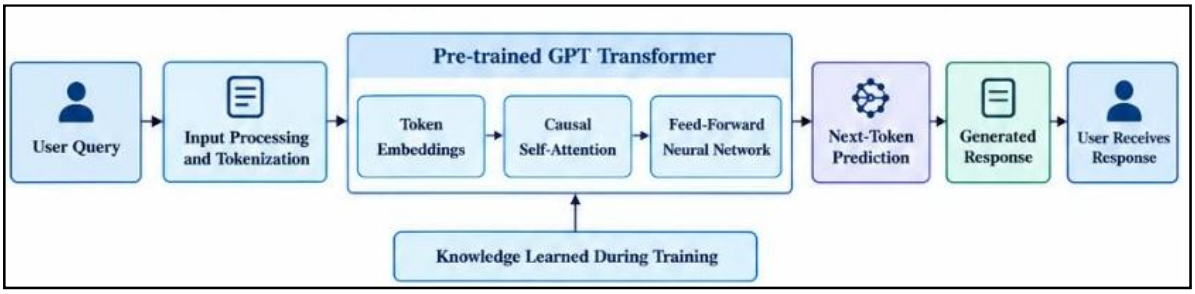


Fig.1: Existing System Architecture of a Conventional GPT-Based System.

2. Tokenization and contextual representations

Text is converted into tokens before being processed by a language model. Tokens may represent complete words, word fragments, punctuation, or other text units. Numerical embeddings represent these tokens within the model.

Tokenization affects computational cost, input length, and multilingual performance. Specialized terminology and languages that are less represented in training data may present additional challenges.

3. Pre-training and post-training

During pre-training, GPT models learn to predict tokens from large collections of text. This process allows them to acquire patterns associated with language, writing styles, factual associations, and common forms of reasoning.

Post-training methods, including supervised fine-tuning and preference-based optimization, can improve instruction following and response quality. However, these methods do not eliminate hallucinations, bias, or unsafe behavior in every context.

The knowledge available within a model is also limited by its training and update processes. A model may not know recent events or organization-specific information unless such information is provided through an appropriate external mechanism.

4. Prompt engineering and context engineering

Prompt engineering involves expressing a task clearly, specifying constraints, and providing examples when useful. Context engineering extends this idea by organizing the information, instructions, tools, and retrieved documents available to the model.

For example, a research assistant may receive a question together with relevant article abstracts, publication dates, and citation details. This context can improve the relevance of its answer, but it does not guarantee that the model will interpret every source correctly.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

5. Retrieval-augmented generation

Retrieval-augmented generation (RAG) connects a language model with an external information-retrieval system. The system retrieves relevant documents and provides them as context for generating a response.

A typical RAG pipeline contains document collection, indexing, retrieval, context selection, generation, and source presentation.

RAG can make it easier to use updated information and private organizational documents without retraining the entire model. However, its effectiveness depends on retrieval quality, source authority, document freshness, and the model's ability to use the retrieved evidence faithfully.

A system may retrieve an incorrect document, overlook contradictory evidence, or cite a source that does not support its claim. Therefore, retrieval must be combined with evidence verification and appropriate uncertainty handling.

6. Fine-tuning and domain adaptation

Fine-tuning adapts a model to a particular task, domain, or response format. Domain-specific adaptation may improve terminology, consistency, and task performance when suitable training data are available.

However, adaptation introduces additional responsibilities, including dataset quality assessment, privacy protection, bias testing, and version management. Fine-tuning should be selected when its expected benefits justify its cost and maintenance requirements.

7. Multimodal systems and external tools

Modern generative AI systems increasingly combine language processing with images, audio, video, and external tools. These capabilities enable broader forms of assistance, such as interpreting visual information or retrieving data from approved services.

Tool-connected systems introduce additional security risks. Malicious instructions embedded in retrieved content, excessive permissions, and incorrectly executed actions can compromise reliability. Secure deployment therefore requires tool restrictions, permission checks, output validation, and human confirmation for consequential actions.

IV. CRITICAL REVIEW OF SELECTED RESEARCH FROM 2024–2026

1. Research published in 2024

Yenduri et al. [2] published a comprehensive review of GPT technologies, enabling methods, applications, challenges, and future directions in *IEEE Access*. Their study provides a broad foundation for understanding GPT architectures, training approaches, and practical use cases.

The paper is important because it connects technical developments with the opportunities and risks of adopting GPT across different domains.

Critical interpretation: A broad technology review provides a useful overview, but the capabilities of generative AI change rapidly. General descriptions of model performance cannot establish that a particular application is reliable under its own operating conditions.

Relevance to the present study: This research establishes the technological background while motivating a more focused investigation of evidence quality, uncertainty, and system-level reliability.

2. Research published in 2025

Saleh et al. [3] examined the evaluation, efficiency, applications, and future directions of large language models in *Frontiers in Computer Science*. Their study is relevant to the question of how models should be assessed beyond their ability to generate fluent text.

Evaluation is particularly important because a model that performs well on one benchmark may produce weaker results on unfamiliar tasks or under different deployment conditions.

Brown et al. [4] published a systematic literature review of retrieval-augmented generation in *Big Data and Cognitive Computing*. Their review examined empirical research on RAG methods, architectures, evaluation practices, and limitations. It provides important context for understanding how external evidence can support language-model outputs while introducing additional retrieval and security challenges.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 7, October 2026 Impact Factor: 8.2

Mienye et al. [5] examined foundational architectures, trends, applications, and ethical considerations in large language models. Their work supports a broader view of language-model development that considers technical design and societal implications together.

Critical interpretation: These studies reinforce the importance of evaluation and external evidence. However, retrieval, generation, and verification should not be treated as interchangeable processes. A system may retrieve relevant documents but still produce a misleading summary, omit contradictory evidence, or provide unsupported citations.

Relevance to the present study: The proposed framework combines evidence retrieval with explicit claim verification, risk assessment, and human oversight.

3. Research published in 2026

Jiang et al. [6] published a survey of large language models in *Frontiers of Computer Science*. The study examines pre-training, post-training, model utilization, and evaluation, providing a broader view of the language-model lifecycle.

Abbasi et al. [7] reviewed ChatGPT applications and evaluation approaches across domains in *Expert Systems*. Their work highlights the variation in evaluation methods and the difficulty of drawing consistent conclusions when studies use different datasets, metrics, and deployment conditions.

Faridi et al. [8] examined the evolution, architectures, applications, and challenges of retrieval-augmented generation in *WIREs Data Mining and Knowledge Discovery*. Their review provides additional context for understanding how retrieval systems are developing beyond basic retrieve-then-generate pipelines.

Critical interpretation: The reviewed research suggests that the reliability of GPT applications depends on the complete system, not merely the language model. Retrieval, context management, verification, tool use, and evaluation can introduce distinct failure points.

Relevance to the present study: The proposed framework treats trustworthy generation as a coordinated process in which evidence quality, model output, risk controls, and human review are evaluated together.

V. APPLICATIONS OF GPT-BASED SYSTEMS

1. Education

GPT can generate explanations, practice questions, summaries, and personalized learning support. It can also help teachers prepare learning materials and provide feedback on student writing.

However, incorrect explanations may mislead students, particularly when they lack the expertise to evaluate the response. Educational applications should encourage independent reasoning, protect student information, and provide access to reliable learning resources.

2. Healthcare

GPT can assist with medical-document summarization, administrative writing, general health-information explanations, and selected clinical documentation tasks.

Its outputs must be carefully validated because errors may affect patient safety. Medical applications require appropriate evidence, privacy safeguards, professional oversight, and domain-specific testing. GPT should not be treated as a replacement for qualified healthcare professionals.

3. Software engineering

GPT can support code generation, debugging, documentation, testing, and explanations of programming concepts. It may reduce the effort required for repetitive development tasks and help beginners understand unfamiliar code.

Nevertheless, generated code can contain security vulnerabilities, logical errors, or inappropriate dependencies. Testing, code review, and secure development practices remain necessary.

4. Business and customer service

Organizations can use GPT to draft communications, summarize reports, classify customer requests, and retrieve information from internal documents.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 7, October 2026 Impact Factor: 8.2

The reliability of these applications depends on the quality of their data, the access permissions assigned to users, and the availability of procedures for escalating difficult cases. Confidential business information must be handled according to appropriate security and privacy policies.

5. Scientific research

GPT can help researchers organize literature, explain technical terminology, develop programming scripts, and improve the clarity of academic writing.

However, generated references may be incorrect, and summaries may omit important methodological limitations. Researchers must verify original publications and disclose AI assistance when required by institutional or publisher policies.

6. Accessibility and multilingual communication

GPT can support translation, text simplification, conversational interfaces, and alternative forms of communication. These capabilities may improve access to information for users with different linguistic and accessibility needs.

Performance should be evaluated across languages, dialects, and user groups. Strong results in one language cannot automatically be generalized to another.

7. Public services and smart-city applications

GPT may assist citizens in navigating public information, understanding administrative procedures, and accessing information from authorized public-service systems.

Applications involving public benefits, legal rights, transport safety, or essential services require strong source controls and clear accountability. Generated explanations must not be confused with official decisions or verified government information.

VI. RESEARCH GAPS

The comparative review identifies six major areas for further investigation.

Gap 1: Fragmented evaluation. Factual accuracy, source quality, latency, cost, and safety are frequently assessed using different methods. A unified evaluation framework could make comparisons more meaningful.

Gap 2: Incomplete claim verification. Retrieval systems can supply relevant documents without ensuring that every important generated statement is supported by those documents.

Gap 3: Limited uncertainty handling. Systems need better procedures for recognizing insufficient evidence, requesting clarification, and abstaining from unsupported conclusions.

Gap 4: Insufficient risk-sensitive control. The level of verification should reflect the potential consequences of an error. A casual writing assistant and a system supporting medical decisions should not necessarily follow identical review procedures.

Gap 5: Limited end-to-end security assessment. Evaluating a language model alone may overlook vulnerabilities introduced by document retrieval, external tools, and automated actions.

Gap 6: Inconsistent long-term monitoring. Models, source documents, and user requirements change over time. Systems need procedures for identifying and responding to declining performance.

These gaps motivate the proposed Human-Centred Evidence-Grounded GPT Framework.

VII. PROPOSED HUMAN-CENTRED EVIDENCE-GROUNDED GPT FRAMEWORK

1. Framework overview

This paper proposes the **Human-Centred Evidence-Grounded GPT Framework (HEG-GPT)** to connect reliable information retrieval, generation, verification, and human oversight within a single application workflow.

The framework is a conceptual contribution developed from the review. It has not yet been implemented or experimentally validated in this paper.

Its central idea is that trustworthiness should be treated as a property of the complete application rather than as a characteristic of the language model alone.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

2. Proposed architecture

The framework contains six stages:

Stage 1: User intent and context analysis

The system identifies the user's objective, constraints, domain, and sensitivity level. Ambiguous requests are clarified when necessary.

Stage 2: Trusted evidence retrieval

When external information is required, the system retrieves documents from approved sources. The evidence is assessed for relevance, authority, freshness, and access permissions.

Stage 3: GPT response generation

The model generates a response using the user request and the selected evidence. The system aims to distinguish sourced statements from explanations, assumptions, and recommendations.

Stage 4: Claim-level verification

Important factual claims are checked against the retrieved sources. Unsupported claims, conflicting evidence, and missing citations are flagged for revision or further investigation.

Stage 5: Risk assessment and human oversight

The system determines whether the response requires additional review based on the task's risk level. Consequential actions require appropriate authorization or human confirmation.

Stage 6: Monitoring and improvement

The system records appropriate evaluation results and authorized feedback. These records can support improvements in retrieval, verification, security, and response quality.

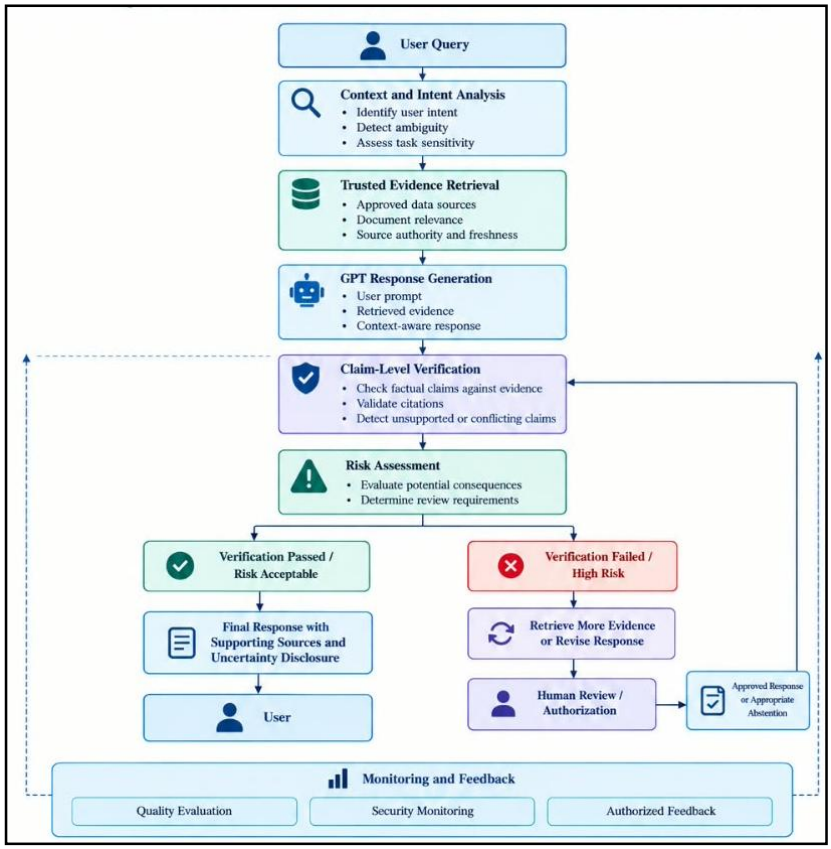


Fig. 2: Proposed Human-Centred Evidence-Grounded GPT Framework (HEG-GPT)

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

3. Architecture flow

The proposed workflow can be represented as:

User Query → Context Analysis → Trusted Evidence Retrieval → GPT Generation → Claim Verification → Risk Assessment → Response or Human Review → Monitoring and Feedback.

The workflow is iterative. If verification fails, the system may retrieve additional evidence, revise the response, ask a clarification question, or abstain from providing a definitive answer.

4. Key design principles

The framework is based on five principles:

1. **Evidence-based responses:** Important factual statements should be supported by relevant and credible sources.
2. **Explicit uncertainty:** The system should communicate when evidence is insufficient or contradictory.
3. **Risk-sensitive verification:** More consequential tasks require stronger checking and review.
4. **Controlled access:** External tools and documents should be accessible only with appropriate permissions.
5. **Human accountability:** AI assistance should not remove the responsibility of the organization or professional deploying the system.

5. Proposed evaluation metrics

Future experiments should assess the framework using predefined metrics.

Evaluation dimension	Suggested measure	Purpose
Factual reliability	Supported-claim rate	Measures factual grounding
Citation integrity	Citation-to-claim support rate	Tests whether cited sources support claims
Retrieval quality	Precision and recall at selected retrieval depths	Evaluates evidence retrieval
Uncertainty handling	Appropriate abstention rate	Tests behavior when evidence is insufficient
Task performance	Task-specific accuracy or completion rate	Measures practical usefulness
Efficiency	Response latency and computational cost	Evaluates operating requirements
Security	Results of predefined attack and access-control tests	Assesses system protection
Human oversight	Compliance with required review procedures	Evaluates accountability

Each metric must be defined precisely before experimentation. For example, the supported-claim rate should use a documented claim-sampling and verification procedure rather than subjective impressions alone.

6. Proposed experimental design

The framework can be evaluated by comparing three configurations:

Baseline A: GPT generation using the user prompt without external retrieval.

Baseline B: GPT with retrieval-augmented generation.

Proposed HEG-GPT: GPT with retrieval, claim verification, risk assessment, and defined human-review procedures.

All configurations should use the same task set, comparable model settings, and consistent evaluation criteria wherever possible.

The experiment should measure factual support, citation integrity, task completion, latency, cost, and failure behavior. Human evaluators should use predefined criteria, and disagreements should be documented.

The study should also examine whether the proposed safeguards introduce additional latency or cost and whether their benefits differ across low-risk and high-risk tasks.

No performance improvement is claimed at this stage. The experiment is a proposed method for determining whether HEG-GPT provides measurable benefits over the baseline configurations.

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

VIII. FUTURE RESEARCH DIRECTIONS

1. Claim-level factual verification

Future research should investigate whether checking individual claims against source documents reduces unsupported statements more effectively than evaluating complete responses alone.

2. Uncertainty-aware generation

Researchers should examine methods for estimating evidence sufficiency and deciding when a model should qualify an answer, ask for clarification, or abstain.

3. Adaptive retrieval strategies

Retrieval systems could adapt to the task, source reliability, information freshness, and evidence gaps. Evaluation should consider both the relevance of retrieved passages and their contribution to final-answer correctness.

4. Efficient and domain-specific models

Comparative experiments should examine whether smaller specialized models can achieve acceptable reliability at lower cost for narrowly defined tasks.

5. Secure tool-using systems

Research should explore how permission restrictions, action validation, and human confirmation affect the safety of GPT systems that interact with external applications.

6. Multilingual reliability

Future work should evaluate whether evidence-grounded approaches perform consistently across languages, dialects, and cultural contexts.

7. Long-term system monitoring

Longitudinal studies should assess changes in accuracy, retrieval quality, and user behavior after model updates or changes in the underlying information sources.

IX. DISCUSSION

The reviewed literature demonstrates that GPT research extends beyond the development of increasingly capable language models. It also concerns how those models are evaluated, connected to external knowledge, and integrated into real applications.

Yenduri et al. [2] provide a broad foundation for GPT technologies and their challenges. Research in 2025 examines evaluation, efficiency, and retrieval-supported generation [3–5]. Studies from 2026 extend this discussion through broader language-model surveys, cross-domain evaluation, and evolving retrieval architectures [6–8].

A common implication is that the quality of a deployed GPT application cannot be inferred from language fluency alone. Its performance also depends on the evidence available to it, the appropriateness of its response, and the controls governing its use.

The proposed HEG-GPT framework translates this observation into a practical architecture. It places claim verification and uncertainty handling between response generation and final delivery, while making risk assessment and human oversight explicit.

However, additional safeguards also introduce trade-offs. Retrieval can increase latency, verification can raise computational costs, and human review may slow the completion of tasks. Automated verification itself can produce false positives or overlook subtle errors.

Therefore, the framework should be tested against suitable baselines rather than assumed to be superior. Its effectiveness will depend on the quality of its implementation, the domain in which it operates, and the reliability of the evaluation process.

X. CONCLUSION

Generative Pre-trained Transformers have expanded the capabilities of artificial intelligence across education, healthcare, software development, business, scientific research, and public services. Their performance is supported by

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

advances in transformer architectures, pre-training, post-training, retrieval-augmented generation, and multimodal processing.

Nevertheless, hallucinations, weak evidence attribution, bias, privacy concerns, security vulnerabilities, and computational costs continue to complicate responsible deployment. Recent research from 2024–2026 highlights the importance of evaluating generative AI through technical, operational, and application-specific perspectives.

This paper proposes the Human-Centred Evidence-Grounded GPT Framework, which combines context analysis, trusted evidence retrieval, GPT generation, claim verification, risk assessment, human oversight, and monitoring. Its primary contribution is a conceptual approach for connecting evidence quality and uncertainty handling with the complete application workflow.

Future research should implement the framework and evaluate it against baseline GPT and RAG systems using consistent tasks, predefined metrics, and reproducible experimental procedures. Such studies would determine whether the proposed approach improves factual reliability and accountability at an acceptable computational and operational cost.

The future of GPT should not be judged solely by its ability to produce fluent responses. It should also be assessed by whether its information can be verified, its uncertainty communicated, its actions controlled, and its outputs used responsibly.

Acknowledgement

I would like to express my sincere gratitude to everyone who supported and encouraged me during the preparation of this research paper, titled **"Rethinking GPT Intelligence: Towards Evidence-Grounded and Trustworthy Generative AI."**

I am especially thankful to my project guide and faculty members of the Department of Computer Engineering for their valuable guidance, suggestions, and encouragement throughout this work. Their support helped me develop a better understanding of Generative Pre-trained Transformers (GPT), trustworthy artificial intelligence, and the challenges associated with the responsible use of generative AI.

I would also like to acknowledge the researchers and authors whose published studies, scholarly articles, and technical contributions provided the foundation for the literature review and comparative analysis presented in this paper.

Finally, I am grateful to my family and everyone who motivated me to complete this research work. Their encouragement and support have been valuable throughout the process.

I sincerely acknowledge all those who contributed directly or indirectly to the completion of this research paper.

REFERENCES

1. A. Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017. Available: https://arxiv.org/abs/1706.03762.
2. G. Yenduri et al., "GPT (Generative Pre-Trained Transformer)—A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions," *IEEE Access*, vol. 12, pp. 54608–54649, 2024, doi: 10.1109/ACCESS.2024.3389497.
3. Y. Saleh, M. Abu Talib, Q. Nasir, and F. Dakalbab, "Evaluating large language models: a systematic review of efficiency, applications, and future directions," *Frontiers in Computer Science*, vol. 7, Art. no. 1523699, 2025, doi: 10.3389/fcomp.2025.1523699.
4. A. Brown, M. Roman, and B. Devereux, "A Systematic Literature Review of Retrieval-Augmented Generation: Techniques, Metrics, and Challenges," *Big Data and Cognitive Computing*, vol. 9, no. 12, Art. no. 320, 2025, doi: 10.3390/bdcc9120320.
5. I. D. Mienye, N. Jere, G. Obaido, O. O. Ogunruku, E. Esenogho, and coauthors, "Large language models: an overview of foundational architectures, recent trends, and a new taxonomy," *Discover Applied Sciences*, vol. 7, Art. no. 1027, 2025, doi: 10.1007/s42452-025-07668-w.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

6. W. X. Zhao, K. Zhou, J. Li, T. Tang, Z. Dong, Y. Hou, B. Zhang, Y. Min, J. Zhang, P. Liu, X. Wang, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, Y. Hu, J.-Y. Nie, and J.-R. Wen, "A Survey of Large Language Models," *Frontiers of Computer Science*, vol. 20, Art. no. 2012627, 2026, doi: 10.1007/s11704-026-60308-3.
7. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020. Available: https://proceedings.neurips.cc/paper/2020/hash/6b493230-Abstract.html.
8. L. Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback," in *Advances in Neural Information Processing Systems*, vol. 35, 2022. Available: https://proceedings.neurips.cc/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.
9. T. B. Brown et al., "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020. Available: https://arxiv.org/abs/2005.14165.
10. R. Bommasani et al., "On the Opportunities and Risks of Foundation Models," arXiv:2108.07258, 2021. Available: https://arxiv.org/abs/2108.07258.
11. S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022, doi: 10.18653/v1/2022.acl-long.229.
12. L. Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," arXiv:2311.05232, 2023. Available: https://arxiv.org/abs/2311.05232.