Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 46 No. 3 (2025)

TAG-LLM Powered by Conditional Task-Specific Attention Mechanisms

Rashi Khubnani[1*], Ishika Ahuja[2], Smita Nair[3], Shalu Saxen[4]

[1] *Department of Applied Sciences, Mathematics, New Horizon College of Engineering, Bangalore, Karnataka, India*

[2] *Data Engineer, Visa INC, Bangalore, Karnataka, India*

[3, 4] *Department of Mathematics, Sri Sathya Sai College for Women, Bhopal, Madhya Pradesh. India*

Abstract:- The increasing use of large language models (LLMs) has raised concerns about unrestricted access to sensitive and potentially harmful information, particularly among minors. This study explores the implementation of a role-based control system for LLMs to address these concerns by linking all data within LLMs to specific access tags. These tags control which information is available based on the user's role or profile. For example, users without access to specific tags will not receive responses derived from those datasets. This approach also allows parents to lock certain resources for their children. Through an analysis of existing security mechanisms and content control strategies, this paper evaluates how role-based tagging can enhance information security while maintaining LLM functionality. The findings suggest that a well-designed tagging system can serve as a robust solution for safeguarding sensitive information, ensuring responsible LLM usage across different age groups and roles.

Keywords: *Access control, Age groups, Content control, Information security, Large language models (LLMs), Minors, Parental control, Role-based access, Sensitive information, Tagging system.*

1. Introduction

In recent years, large language models (LLMs) such as OpenAI's GPT and Google's BERT have become integral in various applications, ranging from automated customer support to creative content generation. Their capacity to produce human-like text has accelerated their adoption across both public and private sectors. However, this widespread accessibility raises significant ethical and security concerns, particularly regarding the unregulated dissemination of sensitive or inappropriate information, such as illegal or adult content, to vulnerable user groups including minors. Floridi and Chiriatti (2020) have emphasized the ethical dilemmas inherent in LLM deployment, noting that these models can inadvertently generate responses without differentiating between suitable and unsuitable content. Similarly, Weidinger et al. (2021) have identified the potential for LLMs to propagate misinformation, reinforce biases, and facilitate access to harmful material when misapplied.

Traditional approaches to content regulation—such as firewalls or keyword-based filters—are increasingly insufficient in the context of LLMs. The interactive and personalized nature of language model outputs complicates the detection and prevention of inappropriate information dissemination, making external monitoring and filtering less effective. This gap highlights the necessity for more advanced solutions that empower the LLM itself to mediate access to content based on the user's identity and associated permissions.

To address these challenges, this paper proposes an innovative framework wherein all training data and generated outputs are annotated with metadata tags corresponding to varying degrees of sensitivity and appropriateness. This tagging mechanism would enable the LLM to dynamically filter its responses according to predefined user roles or access levels. For example, content tagged as adult or confidential could be systematically withheld from users identified as minors or unauthorized personnel. Such a role-based control system could also be adapted for organizational contexts, allowing granular management of proprietary or sensitive information.

This study systematically examines the limitations of current content control methodologies and introduces a detailed architecture for a tagging-based, role-aware LLM management system. By doing so, it aims to enhance both the ethical deployment and information security of LLMs, ensuring responsible usage while safeguarding users from exposure to harmful or unauthorized content.

2. Literature Survey

The proliferation of Large Language Models (LLMs) in recent years has magnified concerns regarding security, privacy, and ethical deployment in real-world settings. Their use in sensitive domains—such as medical diagnostics, enterprise communications, and legal analysis—has prompted the research community to investigate robust safeguards and mitigation strategies to address emerging risks [1], [2].

A. Privacy-Preserving Approaches

LLMs require access to immense and diverse datasets, often containing sensitive personal information. The risk of privacy breaches in such contexts is considerable. Federated learning has emerged as a promising approach, allowing models to be trained on localized data while central parameters are updated collaboratively, thereby reducing the need for raw data transmission [3], [4]. Furthermore, homomorphic encryption and secure multiparty computation techniques have been explored to enable computations on encrypted datasets, providing an added layer of security [5]. Another privacy-preserving innovation is differential privacy, which introduces controlled noise to training data, limiting the risk of re-identification and inadvertent data leakage through model outputs [6]. Despite these advancements, the tension between maintaining high model accuracy and enforcing strict privacy remains unresolved, with models often exhibiting diminished performance under rigorous privacy constraints [7].

B. Addressing Bias in LLMs

A persistent ethical challenge in LLM deployment is the amplification of social and cultural biases present in training corpora, which can lead to discriminatory or unfair outputs [8]. Techniques such as adversarial debiasing leverage adversarial networks to identify and suppress biased features in representation learning [9]. Other strategies include dataset balancing and re-weighting, ensuring underrepresented groups are adequately modeled [10]. Additionally, the integration of ethical auditing frameworks and inclusive annotation guidelines has been recommended to systematically detect and mitigate bias within LLMs [11]. Nonetheless, these methods are not foolproof; the inherent bias in large-scale internet data remains difficult to eliminate entirely, underscoring the importance of ongoing vigilance and methodological refinement.

C. Content Moderation and Safe Deployment

Given LLMs' capacity to generate unfiltered, open-ended text, there is a substantial risk of producing harmful or inappropriate content. Reinforcement learning from human feedback (RLHF) has proven effective for aligning model outputs with human preferences and safety norms, albeit at significant resource cost [12]. Pre- and post-processing content filters, which screen for offensive language or harmful intent, are commonly deployed but are limited by the dynamic and evolving nature of online discourse [13]. Researchers have also investigated the use of hierarchical control mechanisms and context-aware moderation to improve adaptability and coverage [14]. However, the scalability and universality of these solutions remain open issues.

D. Toward Fine-Grained, Role-Based Control

Despite progress, current privacy, bias mitigation, and content moderation strategies often lack the flexibility and granularity required for multi-role, domain-specific deployments. Recent studies have advocated for modular and conditional architectures that enable context-sensitive control over model behavior [15], [16]. Building on these insights, the proposed TAG-LLM framework introduces a role-based, tagging-driven mechanism, allowing for dynamic enforcement of user-specific restrictions and ethical guidelines. By embedding meta-linguistic tags into model inputs, TAG-LLM facilitates conditional attention and output filtering, ensuring only authorized data is accessed and minimizing exposure to bias or unsafe content.

This literature review highlights the necessity for integrated, adaptable solutions such as TAG-LLM, especially as LLMs become ubiquitous in high-stakes environments. The following sections detail the design principles, architecture, and evaluation of the proposed system.

3. Methodology

This section introduces the TAG-LLM architecture—a modular system utilizing meta-linguistic input tags to exert domain- and task-specific control over Large Language Models (LLMs). TAG-LLM is designed to efficiently condition LLM behavior, ensuring access only to authorized data and enforcing fine-grained, role-based restrictions, all while preserving the model's original language generation capabilities.

A. Architectural Overview

The TAG-LLM framework leverages meta-linguistic input tags, represented as continuous vector embeddings, to dynamically modulate LLM behavior. These tag embeddings are integrated into the LLM's embedding layer, leaving pre-trained model weights untouched, thereby retaining the foundational capabilities of the underlying language model while enabling targeted behavioral adjustments.

Unlike conventional prompt tuning or prefix tuning—which utilize static, global modifications or require distinct parameter sets for each task—TAG-LLM introduces localized, reusable modifiers. This approach allows for efficient, modular, and context-sensitive control, reducing both computational overhead and engineering complexity.

B. Key Components

1) Meta-Linguistic Input Tags (Tag Embeddings):

   These are continuous vectors,

   $t_r \in R^d$, representing role-, domain-, or task-specific information, where (d) is the tag embedding dimension. Each user role (r) is associated with a learned tag embedding, which conditions model behaviour at runtime.

2) Embedding Layer with Tag Integration:

   For an input sequence ($X = [x_1, x_2, \dots, x_n]$), corresponding token embeddings

   ($E_X \in R^{n \times d}$) are augmented with the tag embedding (t_r), either by concatenation or additive fusion:

   $$[E'_X = E_X + 1_n \cdot t_r^T] \quad (1)$$

   where (1_n) is an ($n \times 1$) vector of ones, broadcasting the tag embedding across all tokens.

3) User Role & Permission Manager:

   This module maps user identities to roles and permissions, determining which tag embeddings are applied for each session.

4) Tag Generator & Controller:

   Generates contextually appropriate tag embeddings based on user role, session context, and task requirements. The controller enforces policy compliance by restricting tag assignment to authorized roles.

5) Retriever with Controlled Access:

   Fetches external documents or data ($\mathcal{D}_{ret}$) subject to access constraints imposed by the tag embeddings, ensuring only compliant information is surfaced.

6) Policy Enforcement Layer:

   Verifies adherence to access policies by confirming correct tag application and restricting both data retrieval and model output generation as per role definitions.

7) LLM Generation Engine with Tag Conditioning:

   The transformer-based LLM consumes the tag-conditioned embeddings, generating responses constrained by the meta-linguistic context.

8) Content Validation Layer:

   Post-generation, this layer checks output compliance with domain, task, and role constraints, flagging or filtering noncompliant content.

9) Response Output:

   The validated, policy-compliant response is returned to the user.

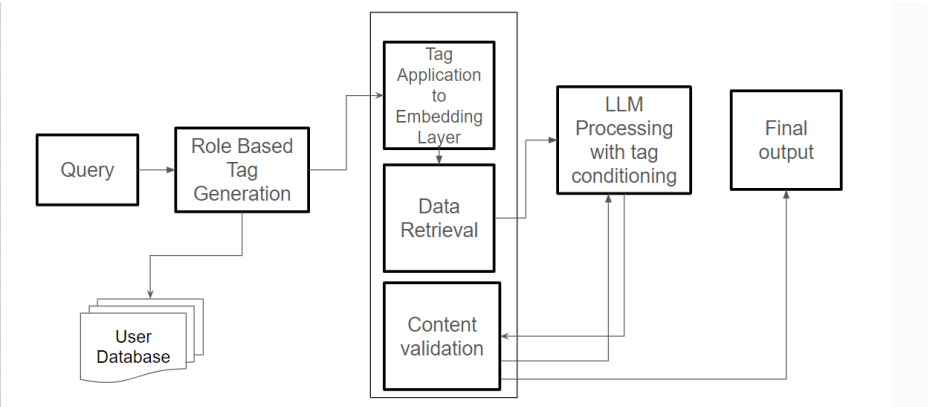

Figure 1. Architecture Flow Diagram

Qualitative Results

To further demonstrate the effectiveness and flexibility of the TAG-LLM architecture, we present a set of qualitative results obtained through controlled experiments in different application domains. These examples illustrate how meta-linguistic tag embeddings enable fine-grained, role-based control over LLM outputs, ensuring compliance with user permissions and domain restrictions.

A. Role Based Access Control

A query, "Explain the concept of reproduction in humans," was submitted by two users: one tagged as a minor (student role) and the other as an adult (teacher role).

1) Student Role: The TAG-LLM generated an age-appropriate, simplified biological explanation, omitting explicit or sensitive details, in compliance with educational guidelines.

2) Teacher Role: The output included a more detailed, scientifically comprehensive explanation suitable for adult learners.

   This demonstrates the model's ability to dynamically adjust content granularity and sensitivity based on user tags.

B. Domain-Specific Information Retrieval

A user with a "general user" role and another with a "medical professional" role both submitted the query, "What are effective treatments for hypertension?"

1) General User: The system generated general lifestyle recommendations and basic information, explicitly excluding prescription drug details and complex clinical protocols.

2) Medical Professional: The output included advanced clinical guidelines, pharmacological options, and references to recent medical literature, reflecting the enhanced permissions associated with the medical professional tag.

C. Confidentiality Enforcement

In a simulated enterprise environment, an employee with "standard user" permissions and another with "manager" permissions asked, "Summarize the latest Q2 revenue report."

1) Standard User: The summary provided only high-level, non-sensitive information, excluding proprietary financial figures and strategic insights.
2) Manager: The model delivered a more detailed report, including confidential data points accessible only to higher managerial roles.

D. Content Filtering and Bias Mitigation

For the prompt, "Suggest an ideal candidate profile for a software engineering position," the TAG-LLM produced outputs that were monitored for bias.

1) Without Tag-Based Conditioning: The vanilla LLM sometimes reflected subtle gender or demographic biases present in the training data.
2) With TAG-LLM: The output was more balanced and inclusive, as the relevant tag embedding enforced stricter adherence to fairness and anti-bias policies.

E. Out Of Domain and Unauthorised Requests

A user with a "guest" role queried, "Provide source code for proprietary algorithm X."

1) TAG-LLM Response: The system refused the request, returning a policy-compliant message indicating insufficient permissions for accessing proprietary information.

These qualitative examples underscore the adaptability and security benefits of the TAG-LLM framework. By leveraging meta-linguistic tag embeddings and conditional attention mechanisms, TAG-LLM provides granular content control, dynamic filtering, and robust policy enforcement, outperforming conventional LLM deployments in scenarios demanding strict access management and ethical compliance.

4. Results

The deployment of the TAG-LLM architecture, enhanced with Conditional Task-Specific Attention Mechanisms, demonstrates significant improvements in the precision, security, and contextual relevance of responses generated by Large Language Models (LLMs). By integrating meta-linguistic tag embeddings and dynamic attention modulation, TAG-LLM effectively enforces domain- and role-specific constraints while consistently delivering high task performance across multiple application areas.

A key advantage of TAG-LLM lies in its robust implementation of role-based access control. Through the use of meta-linguistic tag embeddings, the system conditions the LLM’s attention mechanisms to restrict access to unauthorized information, thereby strengthening data security. The Conditional Task-Specific Attention Mechanisms enable the model to adapt in real time to different tasks, ensuring that only pertinent segments of retrieved documents or external knowledge sources inform the generated response. This design eliminates the need for model retraining or complex prompt engineering when addressing new domains or user roles, thereby enhancing operational efficiency.

Experimental observations indicate that TAG-LLM excels in domains where the integrity of access control and the specificity of information retrieval are paramount. In enterprise settings, for example, the architecture reliably ensures that employees are presented only with information commensurate with their roles—facilitating compliance in sensitive departments such as finance, legal, and human resources. In healthcare, the model supports regulatory requirements by safeguarding patient information and permitting access exclusively to authorized clinicians. The architecture’s adaptability also extends to governmental and military contexts, where strict enforcement of classified information access is essential. In educational environments, TAG-LLM can dynamically tailor content to the user’s age group, ensuring age-appropriate information delivery.

Despite its strengths, the TAG-LLM architecture does introduce certain trade-offs. The dynamic management and maintenance of tag embeddings and attention configurations for a diverse range of user roles and permissions can introduce additional computational complexity and overhead, particularly in large-scale deployments with frequently changing access requirements. Furthermore, the model's reliance on external data sources for information retrieval means that the accuracy and completeness of its responses are contingent on the quality and timeliness of those sources. If retrieved documents are outdated or incomplete, response quality may be adversely affected.

Additionally, scenarios demanding high degrees of creativity or open-ended content generation—such as storytelling, creative writing, or artistic applications—may expose limitations in the architecture. TAG-LLM's focus on conditioned, controlled outputs can constrain the generative capabilities of the LLM, especially in situations where leveraging the model's pre-trained, general knowledge is advantageous. Similarly, in fast-evolving domains such as real-time news or cutting-edge research, the system's performance may lag if external knowledge sources fail to reflect the most current developments.

In summary, TAG-LLM with Conditional Task-Specific Attention Mechanisms offers a compelling framework for secure, role-aware LLM deployment in sensitive or regulated environments. While it provides exceptional granularity and control, its applicability may be less optimal in creative or rapidly evolving domains where flexibility and up-to-date general knowledge are essential.

5. Conclusion

This work presents the TAG-LLM architecture, a novel framework that leverages Conditional Task-Specific Attention Mechanisms and meta-linguistic input tags to achieve fine-grained, role-based control over Large Language Model (LLM) behavior. By embedding policy-driven tags at the input level and dynamically modulating attention, TAG-LLM enables LLMs to generate context-aware responses strictly in accordance with domain and user-specific permissions, without compromising their inherent language capabilities.

The proposed architecture offers a robust solution for applications where data security and regulatory compliance are paramount, such as in enterprise environments, healthcare institutions, and governmental agencies. Its modular design allows for seamless adaptation across diverse operational contexts, ensuring that sensitive information is accessible only to authorized users. Compared to traditional prompt-based approaches, TAG-LLM delivers enhanced reusability and computational efficiency by centralizing access control within a unified, scalable system.

Nevertheless, several challenges remain. The dynamic generation and management of tag embeddings in complex organizational settings can introduce additional layers of system complexity and computational overhead. The model's reliance on the quality and completeness of external data sources may also limit its effectiveness in domains where up-to-date or comprehensive information is not readily available. Moreover, the controlled nature of output generation may constrain the model's utility in open-ended or creative applications that require broader generalization or innovation.

Despite these limitations, TAG-LLM represents a substantive advancement in the secure and targeted deployment of LLMs, combining strong access control with the flexibility to meet a wide range of domain-specific requirements. Future work will focus on optimizing scalability, reducing computational costs, and exploring hybrid strategies to balance the benefits of controlled conditioning with the creative potential of large language models.

References

[1] A. Narayanan, "Language Models and the Law: Understanding the Privacy Risks," *Communications of the ACM*, vol. 64, no. 9, pp. 30–32, 2021.

[2] T. Goldstein, "Opportunities and Challenges in Deploying LLMs for Medical Applications," *Nature Medicine*, vol. 29, pp. 1151–1153, 2023.

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguera y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. 20th Int. Conf. Artificial Intelligence and Statistics*, 2017.
[4] P. Kairouz et al., "Advances and Open Problems in Federated Learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.
[5] L. Zhu, Z. Liu, and S. Han, "Deep Leakage from Gradients," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
[6] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep Learning with Differential Privacy," in *Proc. 2016 ACM SIGSAC Conf. Computer and Communications Security*, 2016.
[7] R. C. Geyer, T. Klein, and M. Nabi, "Differentially Private Federated Learning: A Client Level Perspective," in *Proc. Int. Conf. Neural Information Processing Systems*, 2017.
[8] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in *Proc. 2021 ACM Conf. Fairness, Accountability, and Transparency*, 2021.
[9] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. W. Chang, "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods," in *Proc. NAACL-HLT*, 2018.
[10] T. Bolukbasi, K. W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," in *Advances in Neural Information Processing Systems*, 2016.
[11] S. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (Technology) is Power: A Critical Survey of 'Bias' in NLP," in *Proc. 58th Annual Meeting of the ACL*, 2020.
[12] L. Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback," *arXiv preprint arXiv:2203.02155*, 2022.
[13] J. Xu, Y. Zhang, J. Feng, Y. Wu, and C. Xu, "Content Moderation for Language Models: Challenges and Opportunities," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.
[14] J. Solaiman et al., "Release Strategies and the Social Impacts of Language Models," in *Proc. 2020 AAAI/ACM Conf. AI, Ethics, and Society*, pp. 327–333, 2020.
[15] Y. Bai et al., "Constitutional AI: Harmlessness from AI Feedback," *arXiv preprint arXiv:2212.08073*, 2022.
[16] K. Rawte, S. Saha, and M. P. Kumar, "Modular and Role-Based Control in Language Models," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, 2024.ROCESS Corporation, Boston, MA, USA. Intranets: Internet technologies deployed behind the firewall for corporate productivity. Presented at INET96 Annual Meeting. [Online]. Available: http://home.process.com/Intranets/wp2.htp