Journal of Artificial Intelligence and Technology, 2025, 5, 395-404
https://doi.org/10.37965/jait.2025.0752

ISTP
RESEARCH ARTICLE

Building a Large Language Model for the Human Resource Services in Malaysia

Guan Hong Lai, Darren Xin Lun Chai, and Tong Ming Lim
Faculty of Computing and Information Technology, Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia

(Received 24 February 2025; Revised 02 July 2025; Accepted 26 August 2025; Published online 14 September 2025)

Abstract: This study presents the development of a domain-specific base large language model (LLM) tailored for its application in human resources (HR) management, particularly employee engagement. The model addresses inefficiencies in traditional HR systems, such as inconsistent query resolution on compliance management and policy dissemination. By leveraging LLM's advanced techniques like Rotary Positional Embeddings (RoPE), grouped key-value attention, and Transformer Block enhancements, this research designs and develops an HR-specific base LLM as the first line of HR service for employees in small- and medium-sized enterprises. Data preparation involves cleaning, tokenizing, and training HR-specific datasets, enabling the model to handle nuanced queries with contextual relevance. Through iterative training and evaluation, the Enhanced GPT-2 model has demonstrated significant improvements in learning capability, based on attention weights in the embedding layers, over GPT-2 Small in terms of relevance, consistency, and scalability. Future work focuses on expanding datasets, improving fine-tuning techniques, and integrating retrieval-augmented generation for real-time adaptability.

Keywords: base model; employee engagement; GPT-2; human resources; large language model

I. INTRODUCTION

Human resources (HR) is undergoing a significant transformation driven by technological advancements. Traditional HR systems, which are heavily reliant on manual and database-driven processes to manage employee relations, compliance, and internal policies, frequently face increasing inefficiencies as organizations scale. This has always translated to more HR headcount to accommodate growing needs. These inefficiencies, however, often manifest as delays, human errors, and reduced productivity, particularly in organizations handling complex operations. For instance, the manual processing of each query related to HR can lead to bottlenecks that hinder overall productivity. This underscores the need for innovative, advanced solutions to streamline HR operations [1].

Artificial Intelligence (AI), particularly through the development of large language models (LLMs), has emerged as a promising avenue for addressing these challenges [2]. LLMs show remarkable capabilities in solving text-based tasks such as language understanding, content generation, and contextual analysis. When applied to HR, these capabilities can revolutionize processes by enabling efficient, accurate, and scalable solutions that are designed to meet organizational needs. On many occasions, a domain-specific LLM is generally considered a small language model (SLM), as it is a specialized version of a larger language model to perform well within a specific domain with its unique vocabulary and requirements, often resulting in a smaller model size compared to a general-purpose LLM.

A. BACKGROUND OF STUDY

Recent developments in AI have enabled LLMs to automate complex HR processes and enhance operational efficiency. Traditional AI-powered tools, such as chatbots or sentiment analysis algorithms, are typically limited in scope and unable to provide context-aware responses. For example, a generic chatbot might respond to an inquiry about leave policies with general information, failing to consider specific organizational rules or legal requirements. LLMs, on the other hand, can perform advanced tasks, including content generation, sentiment analysis, and topic classification, with a level of personalization previously unachievable [2].

Despite their potential, generic LLMs often fall short in addressing domain-specific requirements. This limitation arises due to their training on general-purpose datasets, which lack the contextual and specialization richness required for the HR domain. Even though retrieval-augmented generation (RAG) or fine-tuning can address some of these needs, it is dependent on a pretrained LLM that always requires either an API subscription or a self-hosted LLM where its sheer size is an inevitable challenge. Consequently, there is a need for the development of domain-specific models for HR applications that can adapt to HR's unique needs [3].

B. PROBLEM STATEMENT

Current HR systems and generic LLMs face several critical limitations that hinder their effectiveness in real-world applications. First, responses generated by generic LLMs are often irrelevant or inconsistent with user expectations, which undermines their credibility and utility [4]. For instance, a general-purpose LLM might misunderstand the nuances of a question about

Corresponding authors: Guan Hong Lai (e-mail: laigh-wp21@student.tarc.edu.my); Darren Xin Lun Chai (e-mail: darrencxl-wp21@student.tarc.edu.my); Tong Ming Lim (e-mail: limtm@tarc.edu.my).

workplace procedure, leading to incorrect or incomplete guidance.

Second, the opaque nature of LLM architectures hampers transparency and accountability, making it challenging to diagnose and correct errors or biases in generated outputs, leading to potential risks in HR decision-making [5].

Finally, pretrained models struggle to adapt to company-specific requirements, which include unique organizational policies or labor law compliance [3]. This inability to provide tailored solutions often limits the value of generic LLMs in HR applications.

C. OBJECTIVE OF THE STUDY

The primary objective of this study is to develop a domain-specific LLM tailored for HR applications. This involves:

1. Investigating and developing a robust LLM architecture capable of handling diverse HR-related queries with contextual relevance and accuracy.
2. Ensuring consistency, relevance, and human-like fluency in generated responses through advanced natural language generation techniques.
3. Observing and refining the model's outputs to address inaccuracies, improve contextual understanding, and enhance alignment with HR-specific tasks.

D. SCOPE OF THE STUDY

This study focuses on simplifying routine HR tasks, such as answering employee inquiries on company-wide HR benefits, clarifying policies, and ensuring compliance with labor laws. The proposed LLM is designed to specialize in HR-centric functionalities, such as addressing HR-specific queries and tasks, while intentionally excluding deep integration with ancillary systems, such as payroll or recruitment platforms. This decision ensures that the model remains focused, specialized, and optimized for its primary purposes. While future work may explore broader integrations, the current study prioritizes establishing a strong foundation for domain-specific HR applications.

E. SIGNIFICANCE OF STUDY

This research contributes significantly by enhancing HR service efficiency through simplifying repetitive administrative tasks, thereby allowing HR professionals to focus on strategic initiatives. For example, by delegating routine inquiries to an LLM, HR teams can allocate more time to workforce development.

Additionally, the development of a domain-specific LLM improves employee engagement by providing timely, accurate responses to HR queries and ensuring compliance with evolving labor regulations. Furthermore, the scalable nature of the proposed solution positions it as a critical tool for organizations seeking to adapt to future growth and regulatory change. This study also contributes to the broader field of AI by advancing techniques for domain-specific language model development, setting a precedent for similar efforts in other specialized fields.

F. ORGANIZATION OF THE PAPER

The rest of the paper is organized as follows. Section II discusses the related literature review, systematically examining advancements in LLMs and their applications across industries, with particular focus on general-purpose models, domain-specific LLMs, architectural differences, and base model enhancement techniques. Section III presents the methodology and requirements analysis, detailing the key stages of the base modeling process, including requirements analysis, problem identification, data collection and preparation, model design considerations, and transformer block improvements. Section IV covers the base model design, implementation, and evaluation, describing the purpose of the base model, data collection and preparation processes, the Enhanced GPT-2 architecture, model configuration and training procedures, and comprehensive evaluation results with discussion of performance metrics. Section V concludes the paper by summarizing the research contributions and outlining future research directions, including dataset enrichment, scope expansion, and enterprise-level deployment considerations.

II. LITERATURE REVIEW

This section systematically reviews advancements in LLMs and their application across industries, focusing particularly on their relevance to HR. The literature highlights the transformative impact of LLMs and identifies critical challenges such as data diversity, contextual understanding, and domain-specific deployment.

A. GENERAL-PURPOSE MODELS

General-purpose LLMs have revolutionized natural language processing by enabling tasks like text generation, comprehension, and translation. These models, such as GPT, Llama, PaLM, Claude, Gemini, and BERT (summarized in Table 1), leverage advanced architectures and large-scale datasets, offering unmatched versatility across multiple domains.

1. GPT MODELS. GPT models utilize a transformer-based architecture with multi-head attention and autoregressive objectives, excelling in text generation tasks with coherent and contextually accurate outputs [6].

2. LLAMA. Llama introduces innovations such as grouped key-value attention and Rotary Positional Embeddings (RoPE), enhancing efficient long-context processing. This makes it particularly suitable for analyzing extended documents [7].

3. CLAUDE. Claude focuses on ethical AI deployment by incorporating reinforcement learning from human feedback (RLHF), ensuring responses align with human values and provide fairness in decision-making [8].

4. GEMINI. For real-time adaptability, Gemini integrates RAG, enhancing real-time adaptability by retrieving dynamic external information to maintain response accuracy and relevance [9].

5. BERT. BERT, known for its bidirectional transformer architecture, excels at understanding context for comprehension tasks such as classification and sentiment analysis, offering robust contextual embeddings [10].

Despite their capabilities, general-purpose LLMs face challenges such as computational overhead, data bias, and insufficient performance in domain-specific tasks. Addressing these challenges requires architectural enhancements and fine-tuning with domain-specific datasets.

B. DOMAIN-SPECIFIC LLMS

Unlike general-purpose models, domain-specific LLMs, sometimes called SLMs, are tailored to address unique challenges in

Table I. Comparative study on large language models

Author(s)	LLM	Architectures	Strengths
Heidari *et al.*	GPT	Transformer-based with multi-head attention, autoregressive objective, and learned positional embeddings.	Outstanding text generation, scalability across GPT models, and versatility in NLP tasks.
Yin *et al.*	Llama	Grouped key-value attention, Rotary Positional Embeddings (RoPE), and efficient memory utilization.	Effective for long-context processing, reduces memory usage, and is optimized for large datasets.
Hochmair *et al.*	Claude	RLHF-enhanced Transformer.	Ethical AI deployment, aligned responses, and exceptional multi-turn dialog management.
Setzen *et al.*	Gemini	Multimodal transformers integrating retrieval-augmented generation (RAG).	A highly capable multimodal model with real-time adaptability for maintaining accuracy and relevance.
Zalte *et al.*	BERT	Bidirectional transformer architecture with self-attention and masked language modeling objective.	Superior contextual understanding, effective for classification and sentiment analysis tasks.

Table II. Comparative study on LLMs across industries

Author(s)	Industry	Applications
Budhwar *et al.* [11]	Human Resources	Employee queries (leave policies, benefits, workplace guidelines), policy dissemination, compliance support, onboarding, and engagement.
Ma *et al.* [12]	Customer Service	Product troubleshooting, order tracking, sentiment analysis, knowledge management, virtual assistants.
Lee *et al.* [13]	Education	Personalized learning, content generation, grading, student engagement, language learning, and resources for students with disabilities.
Ng *et al.* [14]	Manufacturing	Predictive maintenance, supply chain management, document processing, safety guidelines, and compliance reports.

particular fields. These models overcome the limitations of general-purpose LLMs by incorporating domain-specific knowledge and custom datasets. These models are fine-tuned to understand the intricate language and context of specific fields, ensuring higher accuracy and relevance in their applications (see Table II).

1. HR DOMAIN APPLICATIONS. In the HR domain, domain-specific LLMs support tasks such as

- Responding to employee queries on leave policies, benefits, and regulations.
- Streamlining the dissemination of organizational policies.
- Ensuring compliance with regulatory frameworks [11].

2. OTHER INDUSTRIES.

- Customer Service: LLMs enhance product troubleshooting, order tracking, and sentiment analysis, leading to improved customer experience [12].
- Education: Education-focused LLMs support personalized learning, automated grading, and assistive tools for students with disabilities [13].
- Manufacturing: In manufacturing, these models aid in predictive maintenance, supply chain optimization, and regulatory compliance [14].

Despite their advantages, these models require high-quality datasets, frequent updates to adapt to regulatory changes, and significant computational resources for training.

C. ARCHITECTURAL DIFFERENCES IN LLMs

The architectural designs of LLMs, such as decoder-only models, encoder-only models, and encoder–decoder models, significantly influence their capabilities and applications.

Decoder-only models, such as GPT, employ a unidirectional autoregressive architecture where each token prediction depends on preceding tokens. Stacked transformer decoder blocks with self-attention and feedforward layers make these models highly effective for text generation. However, the unidirectional nature limits bidirectional context utilization, posing challenges for tasks requiring full input understanding [15].

Encoder-only models, such as BERT, use bidirectional transformer encoder layers to capture token relationships across the entire input. This design excels in tasks like classification and sentiment analysis but lacks generative capabilities, limiting its use in tasks requiring text creation [10].

Encoder–decoder models, like T5, combine bidirectional encoding with autoregressive decoding, making them versatile for tasks requiring both comprehension and generation, such as translation or summarization. Despite their adaptability, dual architecture increases computational demands, affecting scalability in resource-constrained settings [16].

D. BASE MODEL ENHANCEMENT TECHNIQUES

Enhancing LLMs for HR-specific tasks involves techniques like prompt engineering, fine-tuning, and RAG. These methods improve contextual relevance, accuracy, and scalability.

1. PROMPT ENGINEERING. This technique structures input queries to guide model responses by specifying key criteria. Properly crafted prompts enhance performance for HR queries but require expertise in prompt design [5].

2. FINE-TUNING. Fine-tuning adapts models to HR contexts by retraining them on domain-specific datasets, improving accuracy in complex scenarios like labor law inquiries. However, it demands high-quality data and substantial computational resources [17].

3. RETRIEVAL-AUGMENTED GENERATION (RAG). RAG integrates LLMs with external knowledge bases for real-time information retrieval. It is particularly useful for dynamic HR contexts, like policy updates, but adds system complexity and requires robust data management [18].

III. METHODOLOGY AND REQUIREMENTS ANALYSIS

This section presents the methodological framework for developing a domain-specific LLM for HR services from scratch. Key stages of the modeling process are illustrated in Fig. 1, with each stage addressing unique challenges and requirements.

A. REQUIREMENTS ANALYSIS AND STUDY

The requirements analysis focuses on addressing critical HR-specific challenges, such as inconsistent query resolution, compliance management, and policy dissemination, emphasizing the need for an efficient, scalable, and contextually accurate LLM. Existing HR systems often struggle to process complex HR language, leading to unreliable and inconsistent responses that can hinder organizational decision-making. Traditional HR chatbots and automation tools rely on keyword-based approaches, lacking the contextual understanding needed to interpret HR queries effectively. Additionally, the vast and evolving nature of HR regulations necessitates a model that can continuously learn and adapt, ensuring its responses remain compliant with legal and organizational policies. Scalability is another crucial factor, as the model needs to handle diverse queries from multiple users across various HR functions without compromising speed or accuracy. Through these analyses, the core requirements of scalability, accuracy, and adaptability were established to align with real-world HR needs, ensuring that the LLM provides reliable, well-structured, and legally sound responses [19].

The study also explores leading LLMs such as GPT-2, GPT-3, and Llama, evaluating their strengths and limitations in domain-specific applications. While these models demonstrate impressive capabilities in general-purpose language tasks, they often struggle with HR-specific scenarios due to the lack of domain-adapted training data and optimization techniques. To bridge this gap, advanced methodologies such as RoPE, grouped key-value attention, and RMSNorm were identified as essential solutions for improving the model's scalability and efficiency. RoPE enhances positional encoding to improve long-form document comprehension, while grouped key-value attention optimizes memory usage, making the model more effective in handling lengthy HR policies and legal documents [20]. These insights guided the integration of advanced techniques for HR applications, addressing domain-specific gaps.

B. PROBLEM AND OBJECTIVE IDENTIFICATION

General-purpose LLMs struggle with HR-specific tasks due to several limitations, including insufficient contextual understanding, reliance on broad datasets, and inadequate positional encoding mechanisms. HR queries often require precise and legally compliant responses, particularly in areas such as employee rights, workplace policies, and labor laws. However, existing models are primarily trained on general internet data, making them prone to providing inaccurate or legally unsound information.

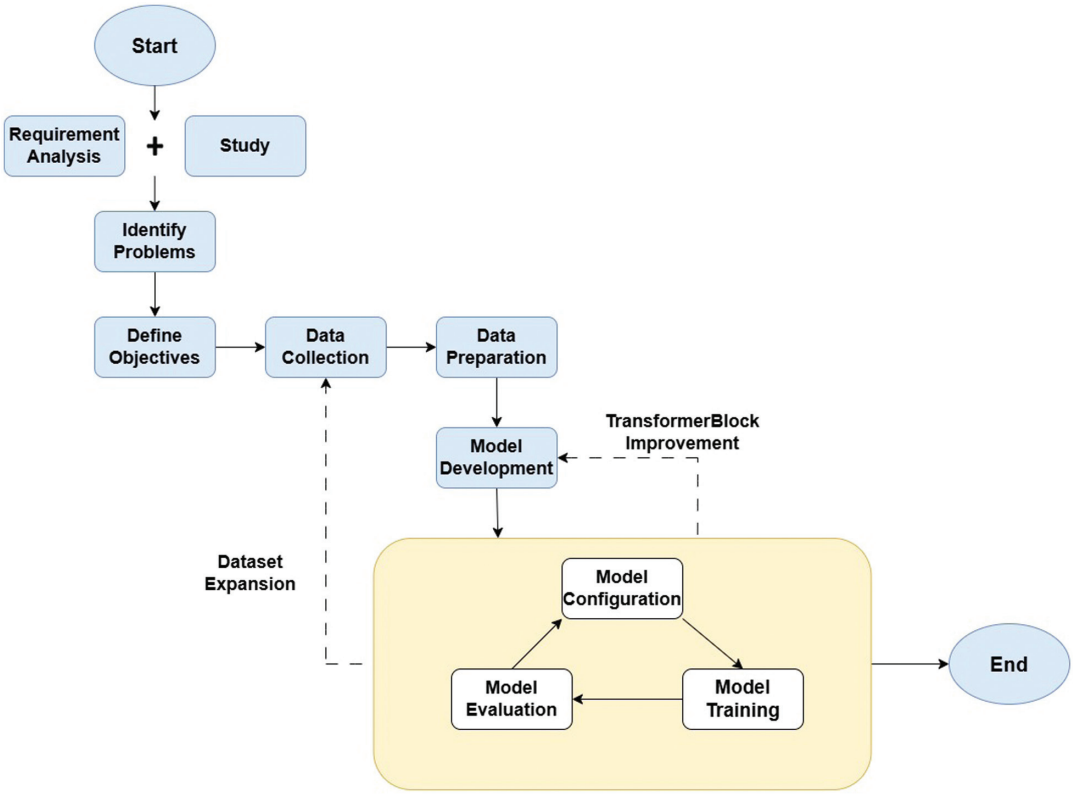

Fig. 1. Key stages of the base modeling process.

Additionally, the processing of long-form documents, such as HR policy manuals and employment contracts, is challenging for standard LLMs, which typically exhibit positional encoding limitations that hinder their ability to maintain coherence over lengthy inputs. These deficiencies highlight the pressing need for a domain-specific LLM tailored to HR operations, ensuring consistency, relevance, and compliance with industry standards and legal frameworks.

The primary objective of this project is to develop an HR-specific LLM capable of delivering accurate, consistent, and contextually relevant responses across various HR domains. The model is designed to specialize in key HR functions such as compliance management, policy interpretation, and employee engagement, providing businesses with a robust AI assistant that enhances HR decision-making. Unlike general-purpose LLMs, this specialized model is structured to process HR-specific jargon, evolving labor laws, and organizational policies, ensuring high precision and ethical compliance. The iterative development approach emphasizes continuous refinement and adaptability, allowing the model to be updated dynamically with new regulations and emerging HR trends. By prioritizing domain relevance, the project aims to create an intelligent HR assistant that streamlines operations, improves efficiency, and supports HR professionals in making informed decisions.

C. DATA COLLECTION

The dataset comprises HR policies, legal documents, training manuals, and employee handbooks sourced from government portals, corporate websites, and legal databases. These documents provide a comprehensive foundation for training the model on policy interpretation, compliance management, and employee engagement scenarios. Data diversity ensures that the model captures various HR contexts, making it adaptable to different organizational needs.

The data collection process emphasizes accuracy and relevance, ensuring that the model remains legally compliant and industry-specific. By integrating case studies and real-world HR applications, the model is designed to understand employee relations, workplace ethics, and regulatory compliance. A dynamic update mechanism allows for the seamless incorporation of new laws and policies, ensuring that the HR LLM remains current and effective in responding to evolving industry requirements.

D. DATA PREPARATION AND PREPROCESSING

Raw HR data undergoes extensive preprocessing to ensure structured training. This involves data cleaning, normalization, and tokenization, with a special focus on preserving domain-specific terminology. Techniques such as bilingual translation, named entity recognition (NER), and stopword removal enhance the model's ability to process HR-related content. HR-specific terms and policies are retained to ensure contextual accuracy.

Additionally, sentence segmentation and document structuring are implemented to help the model process long-form HR documents effectively. The text is lemmatized and standardized, improving readability and interpretability for training. A multi-stage preprocessing pipeline ensures that only relevant and high-quality HR data is included in the model training set. These improvements optimize model performance, enabling it to generate HR-specific responses accurately while maintaining legal and ethical integrity.

E. MODEL DESIGN AND DEVELOPMENT CONSIDERATIONS

The HR-specific LLM is developed by enhancing the GPT-2 Small architecture with Llama-inspired techniques to improve scalability, efficiency, and contextual understanding. One of the key modifications is the integration of RoPE, which significantly improves the model's ability to handle long-form HR documents without degradation in accuracy. Additionally, grouped key-value attention is incorporated to optimize memory usage, enabling the model to process extensive HR policies, employment laws, and organizational guidelines without excessive computational overhead. These improvements make the model more efficient, scalable, and suitable for HR applications requiring detailed and contextually accurate responses.

The Transformer Block enhancements play a crucial role in addressing HR-specific challenges. By fine-tuning attention mechanisms and optimizing token representations, the model improves its ability to maintain coherence across lengthy HR policies and employee manuals. This ensures that queries related to policy interpretation, compliance regulations, and HR best practices are met with precise, well-structured responses. The model's design emphasizes domain specialization, ensuring that it can comprehend, analyze, and generate HR-related content with superior accuracy compared to generic LLMs.

F. MODEL CONFIGURATION, TRAINING, AND EVALUATION IN LOOP

A structured and iterative training approach is employed to continuously refine the model's performance. Training involves hyperparameter tuning, domain-specific dataset fine-tuning, and multiple evaluation cycles to enhance accuracy and contextual relevance. Key hyperparameters, such as learning rate, batch size, and attention mechanisms, are optimized to balance performance, speed, and computational efficiency. The model is trained using RLHF, allowing real-time adjustments based on user interactions and expert evaluations.

Evaluation metrics such as consistency, relevance, and human-likeness are used to assess model quality. Iterative feedback loops ensure that the model adapts dynamically to new HR trends and evolving regulations, keeping responses aligned with real-world HR requirements. Through continuous improvement, the HR-specific LLM is refined to deliver highly reliable, scalable, and ethically compliant solutions, making it a valuable tool for modern HR professionals.

G. TRANSFORMER BLOCK IMPROVEMENT AND DATA EXPANSION

To enhance the model's scalability and contextual comprehension, Transformer Block refinements are implemented, focusing on optimizing attention mechanisms, positional embeddings, and token representations. One of the key improvements is the integration of RoPE, which significantly enhances the model's ability to handle long-form HR documents while maintaining positional coherence. Unlike traditional positional encodings, RoPE enables the model to preserve the relative position of tokens over extended sequences, making it ideal for processing complex policy documents, legal contracts, and HR manuals. Additionally, grouped key-value attention is introduced to optimize memory usage and improve retrieval efficiency, allowing the model to dynamically

adjust its focus based on query complexity and document length. These refinements reduce computational overhead and ensure that responses remain precise, contextually rich, and legally accurate, even when handling multi-paragraph HR queries.

Beyond architectural improvements, data expansion plays a crucial role in ensuring the model remains adaptive to evolving HR industry needs. The dataset is regularly updated with new HR documents, labor law amendments, and workplace compliance policies, ensuring that the model stays aligned with real-time regulatory changes. By incorporating emerging HR trends, such as remote work policies and DEI (Diversity, Equity, and Inclusion) initiatives, and AI-driven HR automation, the model is fine-tuned to address modern workplace challenges. This dynamic dataset update mechanism allows the model to continuously learn, adapt, and evolve, ensuring that HR professionals can rely on it for up-to-date, accurate, and contextually relevant guidance. These combined enhancements make the HR-specific LLM a scalable, efficient, and industry-relevant tool, capable of supporting enterprise HR operations, compliance enforcement, and employee engagement strategies.

IV. BASE MODEL DESIGN, IMPLEMENTATION, AND EVALUATION

This section provides a comprehensive overview of the design, implementation, and evaluation of a domain-specific LLM tailored for HR applications. The following sections detail the purpose of the base model, its architectural enhancements, data preparation process, final configurations, and evaluation results.

A. PURPOSE OF THE BASE MODEL

The base model addresses the limitations of general-purpose LLMs in HR contexts by providing accurate, consistent, and contextually relevant responses. Key HR tasks, such as compliance management, policy interpretation, and employee engagement, require domain-specific understanding, which this model achieves through fine-tuning HR datasets. Fine-tuning allows the model to adapt to specific linguistic patterns and regulatory requirements inherent in HR tasks, enhancing its performance in compliance management and policy interpretation [21].

The objectives of the proposed base model, as shown in Fig. 2, include ensuring human-like interaction with natural and engaging communication and high accuracy for context-aware, actionable answers. A significant focus is placed on mitigating hallucinations, where the model generates plausible but incorrect responses. Techniques such as Dynamic Retrieval Augmentation based on hallucination Detection (DRAD) have been proposed to detect and mitigate hallucinations in LLMs, enhancing their reliability in practical applications [22]. These measures ensure the model's reliability, making it a trusted tool for practical HR applications.

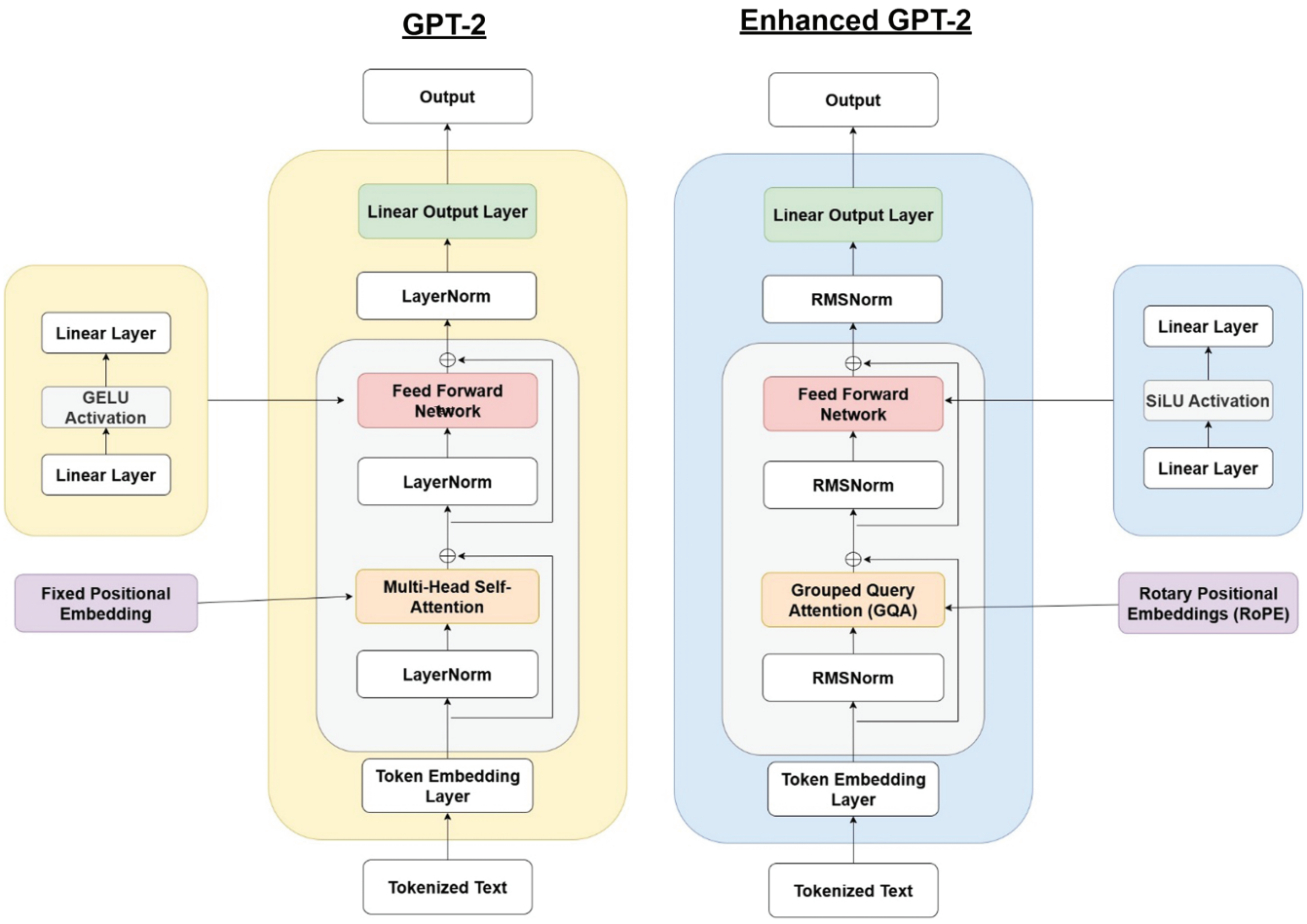

Fig. 2. Architectural design of GPT-2 and Enhanced GPT-2.

Table III. Summary of dataset composition

Dataset type	Files	Token count	Percentage
Legal documents	145	1.7 million	42.5%
HR handbooks and guidelines	78	2.3 million	57.5%

B. DATA COLLECTION AND PREPARATION

The dataset for training the LLM base model is collected from diverse HR-related content to ensure comprehensive coverage of relevant scenarios. Key data sources include Malaysian legal documents, such as employment laws and regulatory guidelines sourced from official government websites, which provide foundational legal compliance knowledge.

Additionally, HR handbooks and company-specific policies contribute practical, day-to-day language relevant to HR operations. Employee training materials and onboarding documents add structured content tailored to role-specific tasks, while industry-specific resources from blogs, books, and journals enrich the dataset with contextual and scenario-based knowledge.

However, the collection process faces challenges, including the need to digitize physical documents using Optical Character Recognition (OCR), which often introduces errors requiring manual correction. Many documents are unstructured, necessitating rigorous cleaning and formatting, while bilingual content is standardized into English using automated translation tools.

Despite these challenges, the resulting dataset is carefully curated to provide a robust foundation for model training. Legal documents account for 42.5% of the total token count, offering detailed insights into compliance and labor laws, while HR handbooks and guidelines make up 57.5%, providing practical language for common HR queries. A detailed summary of the dataset composition is shown in Table III.

C. ENHANCED GPT-2

The base model builds upon the GPT-2 Small architecture, a transformer-based framework renowned for its ability to generate coherent text through autoregressive processes. Key architectural components include:

- **Multi-Head Self-Attention**: Calculates token relationships within sequences for contextual understanding [23].
- **Feedforward Network**: Applies nonlinear transformations to refine token representations.
- **LayerNorm**: Ensures stable training by normalizing input distributions.
- **Fixed Positional Embedding**: Encodes sequential relationships but poses limitations for long-form contexts.

To address the limitations of GPT-2 Small, the Enhanced GPT-2 introduced:

- **Grouped Query Attention (GQA)**: Reduces memory consumption and enhances scalability by grouping query heads and sharing key-value pairs among them [24].
- **Root Mean Square Layer Normalization (RMSNorm)**: Improves training stability and gradient flow by normalizing the root mean square of the input.
- **Sigmoid Linear Unit (SiLU)**: Refines the learning of complex relationships through a smooth, nonlinear activation function.
- **Precomputed Buffer**: Minimizes redundant calculations for efficient inference.

D. BASE MODEL CONFIGURATION & TRAINING

The model training process focuses on configuring and fine-tuning the baseline GPT-2 Small and Enhanced GPT-2 architectures to optimize their performance for HR-specific tasks. Each configuration is iteratively refined to balance scalability, efficiency, and contextual accuracy. The final configurations for both models are summarized in Table IV.

1. GPT-2 SMALL CONFIGURATION. The baseline GPT-2 Small model, featuring six transformer layers and a context length of 512 tokens, is employed as the starting point for this project. Its design includes LayerNorm as the primary normalization mechanism and an independent key-value attention system, which facilitates basic scalability in handling short HR-related queries.

However, the model faces limitations when applied to more complex HR scenarios. The fixed positional encoding mechanism restricts its ability to understand and generate coherent responses for long-context inputs, such as multipage employee handbooks or compliance documents [25]. Additionally, the model's limited hidden dimension hampers its capacity to capture intricate relationships and nuances inherent in HR-specific language.

While the GPT-2 Small configuration provides a reasonable foundation for basic tasks, such as answering short, direct queries, or summarizing brief text, its architectural constraints result in challenges related to scalability and contextual depth. These limitations underscore the need for significant enhancements to address the dynamic and nuanced nature of HR queries effectively.

2. ENHANCED GPT-2 FINAL CONFIGURATION. To overcome the challenges faced by the baseline GPT-2 Small model, the Enhanced GPT-2 configuration is developed with 12 transformer layers, an extended context length of 768 tokens, and advanced architectural improvements inspired by state-of-the-art techniques.

Key enhancements include the integration of RoPE and grouped key-value attention. The RoPE mechanism significantly improves the model's ability to handle long-context inputs by encoding relative positional information directly into the attention mechanism, thereby preserving relationships between tokens across extended sequences [26]. This innovation is particularly beneficial for processing HR documents that require comprehension of hierarchical structures, such as employee policies or multi-turn conversations.

Additionally, the inclusion of precomputed shared buffers for attention masks and RoPE parameters reduces redundant computations, further improving memory efficiency and response times [27]. This is particularly advantageous in enterprise-level applications, where real-time processing and scalability are essential.

Table IV. Key configuration differences of GPT-2 and Enhanced GPT-2

Configuration parameter	GPT-2 small	Enhanced GPT-2
Vocabulary size	50,257	50,257
Embedding dimension	768	768
Context length	512	768
Number of attention heads	6	12
Number of transformer layers	6	12
Hidden dimension	–	3072
Dropout rate	0.2	0.3
Key-value bias	False	True

With its larger hidden dimension of 3072, the model demonstrates a significantly improved ability to capture subtle patterns and relationships within HR-related language, leading to more coherent and contextually relevant responses [28].

By iteratively refining the hyperparameters and introducing architectural improvements, the Enhanced GPT-2 demonstrates superior adaptability, efficiency, and contextual understanding compared to the baseline GPT-2 Small. These configurations underline the advancements made to tailor the model for HR-specific tasks.

E. BASE MODEL EVALUATION AND DISCUSSION

The evaluation and discussion for the base model focus on assessing the performance of GPT-2 Small and Enhanced GPT-2 models. Outputs are evaluated using a star rating system (1–5 stars) across key metrics: relevance, sentence-level consistency, overall consistency, and human-like quality. The maximum rating assigned is four stars to reflect those outputs that still require significant refinement (as shown in Table V).

1. EVALUATION METRICS AND RESULTS. Relevance measures how effectively the generated output addresses the query posed by the user, ensuring that the response is accurate and meaningful. Sentence-level consistency evaluates the logical flow and coherence within individual sentences, ensuring that ideas are articulated and error-free. Overall consistency assesses the logical cohesion and thematic relevance across the entire response, ensuring that the output maintains a steady tone and theme throughout. Human-like quality captures the naturalness of the response in terms of tone, style, and language fluency, determining how closely the generated text resembles professional human communication.

2. DISCUSSION OF RESULTS. The GPT-2 Small model demonstrates reasonable performance in addressing general HR-related queries but struggles significantly with more procedural and complex topics. For instance, its outputs are relevant and well structured at the sentence level but often lack coherence and consistency across the entire response. It occasionally produces irrelevant or incomplete answers, highlighting gaps in the training dataset. While the model's responses often display a human-like tone, they frequently fail to provide detailed procedural steps, revealing their limitations in tackling nuanced and context-sensitive queries.

In comparison, Enhanced GPT-2 exhibits notable improvements across all metrics. It consistently generates relevant, coherent, and structured responses that effectively address queries. Enhanced GPT-2 demonstrates a superior understanding of complex topics, presenting advice comprehensively and with greater accuracy. Its sentence level and overall consistency are significantly better, though occasional repetition and lack of depth in procedural explanations persist. These issues indicate areas for further enhancement, particularly in addressing detailed and context-rich queries.

Table V. Star rating evaluation

Model	GPT-2 small	Enhanced GPT-2
Relevance	★★★☆☆	★★★★☆
Consistency	★★★☆☆	★★★★☆
Overall consistency	★★☆☆☆	★★★☆☆
Human-like	★★★☆☆	★★★★☆

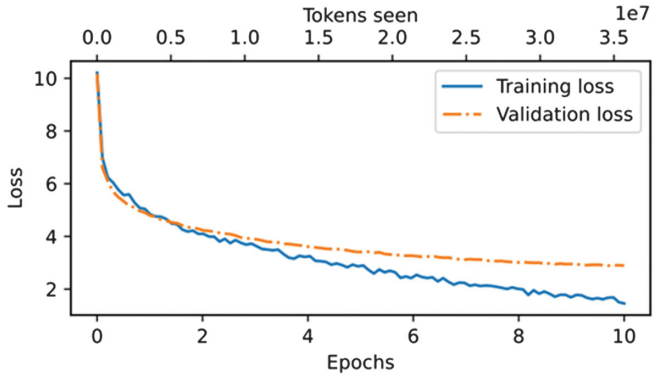

Fig. 3. Training and validation loss trends across Epochs for the base model.

When comparing the two models, Enhanced GPT-2 outperforms GPT-2 Small in all evaluated metrics, particularly in relevance and consistency. The improvements are most evident in queries requiring actionable advice and structured responses. However, both models reveal limitations, especially in handling complex procedural scenarios, suggesting the need for a more diverse dataset and advanced fine-tuning techniques to further improve output quality.

3. LOSS ANALYSIS AND CONVERGENCE. The training and validation loss trends during the Enhanced GPT-2 base model are visualized in Fig. 3, providing insights into the model's performance across 10 epochs.

4. KEY OBSERVATIONS.

1. **Training Loss:** The training loss exhibits a steep decline during the initial epochs, demonstrating effective learning as the model adjusts to the training data. However, in subsequent epochs, the training loss continues to decrease steadily, indicating that the model is increasingly optimizing for the specific patterns in the training dataset.
2. **Validation Loss:** The validation loss decreases during the early epochs, reflecting improved generalization. However, it begins to plateau around the fourth epoch, with minimal improvement in later epochs. This trend suggests that the model's ability to generalize to unseen data reaches its peak early in the training process.
3. **Gap Between Losses:** A noticeable gap between training and validation losses emerges as training progresses. While the training loss continues to decline, the validation loss stabilizes, indicating a divergence in the model's performance on the training versus validation datasets. This gap is a hallmark of overfitting, where the model learns to perform exceedingly well on the training data but struggles to maintain consistency on unseen data.

V. CONCLUSION AND FUTURE STUDY

This research has demonstrated the potential of a domain-specific LLM base model tailored for HR applications. The Enhanced GPT-2 model, developed using HR-specific datasets and iterative evaluation, has shown significant promise in addressing challenges such as improving query handling and enhancing employee experiences. By focusing on creating a robust base model, this study

has laid the groundwork for future advancements in HR-specific AI solutions.

Future research will emphasize several key directions to expand the model's capabilities, generalizability, and scalability. First, enhancing ***prompt engineering*** with detailed instructions and contextual hints will improve the quality of responses, particularly in complex or sensitive HR scenarios. To ensure that outputs are polished and professional, advanced postprocessing pipelines will be explored to refine grammar, structure, and tone.

Dataset enrichment remains a critical priority. The current 4 million token corpus, while a useful starting point, was insufficient to fully capture the diversity and intricacies of HR-related language. Future efforts will focus on substantially expanding the dataset with a broader range of domain-specific content, including company policies, employee handbooks, legal frameworks, training manuals, and case studies. A richer dataset will significantly improve the model's ability to generalize to diverse HR scenarios and emergent workplace challenges.

Additionally, ***the scope of application*** will be broadened to explore the model's utility across a wider range of HR functions, such as talent acquisition, learning and development, performance management, and employee engagement analytics. Integration with RAG will also be pursued to allow real-time access to evolving information, ensuring that model outputs remain contextually relevant and up to date.

Past efforts have targeted scaling the model for ***enterprise-level deployment***, improving adaptability to dynamic organizational contexts, and expanding its use cases. These advancements aim to create a scalable, context-aware, and intelligent framework that supports the digital transformation of HR operations while enhancing overall employee experience.

CONFLICT OF INTEREST STATEMENT

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

[1] K. Sanders, L. J. Song, Z. Wang, and T. C. Bednall, "New frontiers in HR practices and HR processes: evidence from Asia," *Asia Pac. J. Hum. Resour.*, vol. 60, no. 4, pp. 703–720, 2022.

[2] D. Yigci, M. Eryilmaz, A. K. Yetisen, S. Tasoglu, and A. Ozcan, "Large language model-based chatbots in higher education," *Adv. Intell. Syst.*, vol. 7, no. 3, p. 2400429, 2024.

[3] J. O. Krugmann and J. Hartmann, "Sentiment analysis in the age of generative AI," *Springer*, vol. 11, p. 3, 2024.

[4] U. Alkafaween, I. Albluwi, and P. Denny, "Automating autograding: large language models as test suite generators for introductory programming," *J. Comput. Assist. Learn.*, vol. 41, no. 1, p. e13100, 2024.

[5] C. W. Safranek, T. Huang, D. S. Wright, C. X. Wright, V. Socrates, R. B. Sangal, M. Iscoe, D. Chartash, and R. A. Taylor, "Automated HEART score determination via ChatGPT: honing a framework for iterative prompt development," *J. Am. Coll. Emerg. Physicians Open*, vol. 5, no. 2, p. e13133, 2024.

[6] A. Heidari, N. J. Navimipour, S. Zeadally, and V. Chamola, "Everything you wanted to know about ChatGPT: components, capabilities, applications, and opportunities," *Internet Technol. Lett.*, vol. 7, no. 6, p. e530, 2024.

[7] C. Yin, K. Du, Q. Nong, H. Zhang, L. Yang, B. Yan, X. Huang, X. Wang, and X. Zhang, "PowerPulse: power energy chat model with LLaMA model fine-tuned on Chinese and power sector domain knowledge," *Expert Syst.*, vol. 41, no. 3, p. e13513, 2023.

[8] H. H. Hochmair, L. Juhász, and T. Kemp, "Correctness comparison of ChatGPT-4, Gemini, Claude-3, and Copilot for spatial tasks," *Trans. GIS*, vol. 28, no. 7, pp. 2219–2231, 2024.

[9] S. A. Setzen, K. Andreadis, O. Elemento, and A. Rameau, "AI-powered laryngoscopy: exploring the future with Google Gemini," *Laryngoscope*, vol. 135, no. 6, pp. 1851–1853, 2025.

[10] J. Zalte and H. Shah, "Contextual classification of clinical records with bidirectional long short-term memory (Bi-LSTM) and bidirectional encoder representations from transformers (BERT) model," *Comput. Intell.*, vol. 40, no. 4, p. e12692, 2024.

[11] P. Budhwar *et al.*, "Human resource management in the age of generative artificial intelligence: perspectives and research directions on ChatGPT," *Hum. Resour. Manag. J.*, vol. 33, no. 3, pp. 606–659, 2023.

[12] X. Ma, R. Zhao, Y. Liu, C. Deng, and D. Du, "Design of a large language model for improving customer service in telecom operators," *Electron. Lett.*, vol. 60, no. 10, p. e13218, 2024.

[13] J. Lee, Y. Hicke, R. Yu, C. Brooks, and R. F. Kizilcec, "The life cycle of large language models in education: a framework for understanding sources of bias," *Br. J. Educ. Technol.*, vol. 55, no. 5, pp. 1982–2002, 2024.

[14] W. L. Ng, G. L. Goh, G. D. Goh, J. S. J. Ten, and W. Y. Yeong, "Progress and opportunities for machine learning in materials and processes of additive manufacturing," *Adv. Mater.*, vol. 36, no. 34, pp. 1–20, 2024.

[15] A. Lopez-Lira,*The Predictive Edge: Outsmart the Market Using Generative AI and ChatGPT in Financial Forecasting*. Hoboken, NJ, USA: John Wiley & Sons, 2024.

[16] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[17] Z. Huang, P. Peng, F. Lu, and H. Zhang, "An LLM-based method for quality information extraction from web text for crowd-sensing spatiotemporal data," *Trans. GIS*, vol. 29, no. 1, pp. 123–140, 2024.

[18] S. Koga, D. Ono, and A. Obstfeld, "Retrieval-augmented generation versus document-grounded generation: a key distinction in large language models," *J. Pathol. Clin. Res.*, vol. 11, no. 1, pp. 1–10, 2025.

[19] R. A. Sithambaram and F. P. Tajudeen, "Impact of artificial intelligence in human resource management: a qualitative study in the Malaysian context," *Asia Pac. J. Hum. Resour.*, vol. 61, no. 4, pp. 789–807, 2022.

[20] D.-J. Li, Y.-C. Kao, S.-J. Tsai, Y.-M. Bai, T.-C. Yeh, C.-S. Chu, C.-W. Hsu, S.-W. Cheng, T.-W. Hsu, C.-S. Liang, and K.-P. Su, "Comparing the performance of ChatGPT GPT-4, Bard, and Llama-2 in the Taiwan psychiatric licensing examination and in differential diagnosis with multi-center psychiatrists," *Psychiatry Clin. Neurosci.*, vol. 78, no. 6, pp. 321–330, 2024.

[21] D. Endalie, "Fine-tuning BERT models for multiclass Amharic news document categorization," *Complexity*, vol. 2025, no. 1, pp. 1–12, 2025.

[22] W. Su, Y. Tang, Q. Ai, C. Wang, Z. Wu, and Y. Liu, "Mitigating entity-level hallucination in large language models," *arXiv preprint arXiv:2407.09417*, 2024.

[23] W. Yan, B. Zhang, M. Zuo, Q. Zhang, H. Wang, and D. Mao, "AttentionSplice: an interpretable multi-head self-attention based

hybrid deep learning model in splice site prediction," *Chin. J. Electron.*, vol. 31, no. 5, pp. 889–899, 2022.

[24] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai, "GQA: training generalized multi-query transformer models from multi-head checkpoints," *arXiv preprint arXiv:2305.13245*, May 2023.

[25] Y. Chen and J. Yan, "What rotary position embedding can tell us: identifying query and key weights corresponding to basic syntactic or high-level semantic information," in *NeurIPS*, December 2024.

[26] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "RoFormer: enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 552, pp. 1–14, 2023.

[27] J. Yang, B. Hou, W. Wei, Y. Bao, and S. Chang, "KVLink: accelerating large language models via efficient KV cache reuse," *arXiv preprint arXiv:2502.16002*, February 2025.

[28] M. Uzair and N. Jamil, "Effects of hidden layers on the efficiency of neural networks," in *Proc. IEEE Int. Conf. Comput., Math. Eng. Technol. (iCoMET)*, Islamabad, Pakistan, January 2020, pp. 1–6.