Journal of Artificial Intelligence in Information Technology (JAI-IT)
Vol. 1 No. 1 June 2026
E-ISSN: XXXX-XXXX; P-ISSN: XXXX-XXXX, Pages 191-211
DOI: https://doi.org/10.51903/3467pq60

Transformer Based Intelligent Virtual Assistant for Automated IT Helpdesk Resolution: A System Implementation and Comparative Evaluation Study

Liza Putri Pagan*[1], Maya Utami Dewi[1]
Email: liza@stekom.ac.id (1), maya@stekom.ac.id (2)
Orcid: https://orcid.org/0009-0002-1593-778X (1), https://orcid.org/0009-0009-4060-5067 (2)
[1]*Department of Information Technology, Faculty of Academic Studies, Universitas Sains dan Teknologi Komputer, Semarang, Indonesia 50192*
*Corresponding Author

Abstract
Enterprise IT service management environments face mounting operational pressure as the volume and complexity of support requests increasingly exceed the capacity of conventional human operated helpdesk systems. This study addresses that challenge by designing, implementing, and evaluating a transformer based intelligent virtual assistant for automated IT helpdesk resolution within an enterprise ITSM workflow. Three pre-trained transformer architectures BERT, RoBERTa, and DistilBERT were fine tuned on a publicly available IT helpdesk ticket dataset; following preprocessing and removal of duplicate and incomplete records, 4,800 usable annotated instances were retained, spanning five intent categories: hardware failure, software malfunction, network connectivity, access management, and general inquiry. The system incorporates dual task inference combining intent classification with retrieval based response generation, governed by a confidence gated escalation mechanism that routes low confidence predictions to human agents. RoBERTa achieved the highest classification accuracy at 93.6% with a weighted F1-score of 0.934, while DistilBERT reduced inference latency by 45.8% relative to RoBERTa, offering a computationally efficient alternative for latency-constrained deployments. At the system level, the RoBERTa configuration attained a ticket deflection rate of 92.8% under offline evaluation conditions, confirming the operational viability of the proposed architecture for autonomous first-line incident resolution within the scope of the experimental setup reported here. These findings provide practitioners with empirically grounded, multi criteria guidance for transformer model selection in enterprise helpdesk deployment, and contribute a replicable integration architecture that bridges the gap between isolated model evaluation and production-representative ITSM implementation documented in prior literature.

Keywords: *BERT, IT service management, natural language processing, RoBERTa, ticket classification*

I. INTRODUCTION

The accelerating pace of digital transformation across enterprises has fundamentally altered the operational demands placed upon Information Technology (IT) service management functions. As organizations progressively migrate to cloud based infrastructures and adopt distributed architectures, the volume, complexity, and diversity of IT support requests have grown at a rate that renders conventional, human operated helpdesk systems increasingly inadequate [1]. Traditional IT service desk operations rely heavily on manual triaging, which introduces response latency, inconsistent resolution quality, and substantial labor overhead challenges that are particularly acute in large scale enterprise environments operating across multiple time zones [2]. The cumulative cost of IT service inefficiency extends beyond monetary expenditure, directly

Received on March, 2026; Revised on April, 2026; Accepted on April, 2026; Published on June, 2026

impacting organizational productivity, employee experience, and business continuity. Consequently, the automation of IT helpdesk functions has emerged as a strategic priority for enterprise IT governance frameworks, necessitating the exploration of robust, scalable, and intelligent solutions capable of handling natural-language service requests with minimal human intervention.

Natural Language Processing (NLP) has undergone a paradigm shift with the advent of transformer based architectures, most notably the Bidirectional Encoder Representations from Transformers (BERT) model [3], which established a new benchmark for contextual language understanding across a broad spectrum of downstream tasks. Subsequent architectural refinements, including RoBERTa [4] and DistilBERT [5], domain adapted variants, have extended the applicability of transformer models to specialized textual domains such as technical documentation and IT specific communication patterns [6]. The self attention mechanism inherent to these architectures enables the model to capture long range token dependencies and semantic nuances that earlier recurrent models such as LSTM and GRU struggled to represent effectively. Applied to IT helpdesk scenarios, transformer models offer the capacity to interpret ambiguous user queries, classify incident categories with high precision, and generate contextually appropriate resolution responses capabilities that collectively define the functional requirements of an enterprise grade intelligent support system.

A growing body of research has explored the integration of NLP driven virtual assistants within enterprise and IT service environments. Previous studies [1] demonstrated that fine-tuned BERT models applied to IT ticketing corpora achieved statistically significant improvements in intent classification accuracy over conventional machine learning baselines, while [7] developed a transformer-based dialogue management system for enterprise service desks and reported measurable reductions in mean time to resolution (MTTR) following deployment. According to [8] demonstrated that pre-trained language models such as BERT significantly improve matching accuracy in structured data integration tasks, an approach that can be adapted to link IT tickets with known resolution records in knowledge base retrieval systems. A prior study [7] introduced a hybrid architecture combining rule-based escalation logic with neural intent recognition that consistently outperformed purely neural approaches on out of distribution inputs, reinforcing the argument that production-grade deployment requires architectural considerations that extend well beyond isolated model optimization.

Despite the volume of research demonstrating the technical viability of transformer based NLP in service automation contexts, critical gaps persist that limit direct applicability to enterprise IT helpdesk environments. The majority of prior studies evaluate model performance in isolation,

without embedding the NLP component within a fully operational system architecture that reflects production level constraints such as authentication, session management, and integration with ITSM platforms such as ServiceNow or Jira [9]. Comparative evaluations between transformer variants are frequently conducted on generic benchmark datasets that bear limited resemblance to the specialized lexical and syntactic properties of IT incident records [10], while most existing implementations report only academic performance measures rather than operationally meaningful indicators such as deflection rate, first contact resolution improvement, or user satisfaction scores. This confluence of gaps limits the degree to which existing findings can inform enterprise deployment decisions, warranting a study that simultaneously addresses system implementation, domain specific comparative evaluation, and operational performance measurement within a unified research design.

This study aims to design, implement, and evaluate a transformer based intelligent virtual assistant for automated IT helpdesk resolution, with the primary objective of constructing a functional system architecture that integrates a fine tuned transformer model with a structured ITSM workflow. A secondary objective involves a rigorous comparative evaluation of BERT, RoBERTa, and DistilBERT using an IT domain specific incident dataset, providing evidence based model selection guidance for practitioners. The contributions of this research are threefold: technically, it delivers a replicable integration architecture linking NLP components to ITSM platform functions [11]; empirically, it provides domain relevant comparative benchmarking of transformer variants that extends beyond generic evaluation norms; and practically, it incorporates operationally meaningful performance metrics alongside deployment considerations including scalability, data privacy, and human escalation pathways to serve IT managers and enterprise architects responsible for AI adoption governance.

II. RESEARCH METHOD

A. Research Design

This study adopts a quantitative research approach with an experimental design, wherein system development and empirical model evaluation are conducted under controlled conditions to produce measurable, reproducible outcomes. The experimental design is selected on the grounds that it enables direct performance comparison across multiple transformer model variants specifically BERT, RoBERTa, and DistilBERT under identical training, validation, and testing conditions, thereby ensuring that observed performance differences are attributable to architectural distinctions rather than procedural inconsistencies [12], [13]. The study further incorporates a system development component following a structured implementation framework, in which the fine-tuned NLP model is embedded within a functional IT helpdesk

architecture capable of handling both intent classification and automated response generation. This dual-component design combining model-level experimentation with system-level implementation directly addresses the gap identified in the introduction regarding the absence of production-contextualized evaluation in prior transformer-based ITSM research [14], [15].

The overall research workflow is structured into five sequential phases, as illustrated in Figure 1. Phase 1 encompasses dataset acquisition and preprocessing, wherein the publicly available IT ticket dataset is retrieved, cleaned, and partitioned for experimental use. Phase 2 covers model configuration and fine-tuning, applying transfer learning from pre-trained transformer checkpoints to the IT-domain dataset. Phase 3 constitutes system integration, embedding the fine-tuned model within the proposed virtual assistant architecture. Phase 4 involves comparative evaluation, assessing all three model variants against standardized classification and operational metrics. Phase 5 concludes with result analysis and interpretation, wherein findings are contextualized against existing literature and practical deployment implications are drawn.

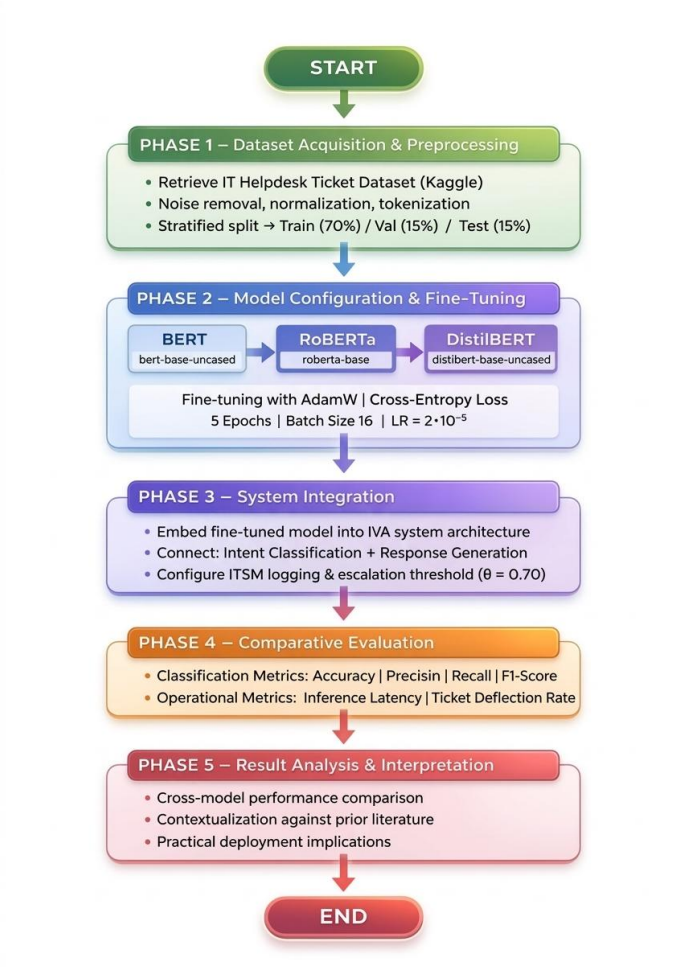

Figure 1. *Research Design Flowchart of the Proposed Study*

B. Dataset

The dataset employed in this study was obtained from a publicly available IT support ticket repository on Kaggle, namely the Customer IT Support - Ticket Dataset [16]. The dataset comprises real-world enterprise IT support tickets, including textual descriptions and associated categorical labels. After data preprocessing, which involved the removal of duplicate entries and records with missing or incomplete annotations, a total of 4,800 valid instances were retained for subsequent analysis and model evaluatio. The dataset contains annotated incident records spanning multiple intent classes including hardware failure, software malfunction, network connectivity, access management, and general inquiry categories that reflect the predominant ticket taxonomy observed in production ITSM deployments [14]. Each record consists of a free-text ticket description submitted by end users, accompanied by a corresponding resolution note generated by IT support agents, providing the dual-label structure necessary to support both the classification and response generation tasks addressed in this study. Prior to model training, the dataset undergoes stratified splitting into training (70%), validation (15%), and testing (15%) subsets to ensure proportional class representation across all experimental partitions, mitigating the risk of class imbalance bias in performance evaluation [17], [18]. The dataset distribution across intent categories is summarized in Table 1.

Table 1. *Dataset Distribution Across Intent Categories*

Intent Category	**Description**	**Proportion**
Hardware Failure	Issues related to physical device malfunction	~20%
Software Malfunction	Application errors and system crashes	~25%
Network Connectivity	Internet, VPN, and LAN-related incidents	~20%
Access Management	Password resets, account lockouts, permissions	~20%
General Inquiry	Non-critical requests and information queries	~15%

C. *Data Preprocessing*

Raw ticket records retrieved from the Kaggle repository require systematic preprocessing before being suitable for transformer model input. The preprocessing pipeline comprises five sequential stages: (1) noise removal, wherein HTML tags, special characters, and non-alphanumeric tokens irrelevant to semantic content are stripped from raw text; (2) case normalization, applying lowercase conversion to reduce vocabulary dimensionality; (3) tokenization using model-specific tokenizers BertTokenizer for BERT and DistilBERT, and RobertaTokenizer for RoBERTa each of which applies WordPiece or Byte-Pair Encoding (BPE) subword segmentation as appropriate to the respective architecture (4) sequence truncation and padding to a maximum token length of

128, a length selected based on the token distribution of the dataset, wherein the majority of ticket descriptions (approximately 95%) contained fewer than 128 tokens after tokenization, ensuring that truncation does not result in meaningful loss of semantic content for the predominant portion of records; and (5) label encoding, converting categorical intent labels into integer indices compatible with the classification head of each model. The preprocessing pipeline is implemented uniformly across all three model variants to ensure that observed performance differences reflect architectural properties rather than input preparation discrepancies.

D. System Architecture

The proposed intelligent virtual assistant system is structured around three functional layers, as illustrated in Figure 2. The input layer receives free-text user queries submitted through a conversational interface and routes them to the NLP processing layer. The NLP processing layer comprises the fine-tuned transformer model, which performs dual-task inference: first, intent classification to determine the incident category of the submitted query; and second, response generation, wherein the classified intent is mapped to a retrieval-based response selected from a curated resolution knowledge base constructed from a curated resolution knowledge base.

The knowledge base is constructed from the training corpus by indexing resolution notes associated with each training record, organized by intent category. At inference time, given a predicted intent class, the system retrieves the most semantically relevant resolution note from the corresponding category partition using cosine similarity over TF-IDF representations. In cases where no candidate response exceeds a minimum relevance threshold, a predefined category-level default response template is returned, ensuring that the system always delivers a response rather than failing silently.

The output layer delivers the generated response to the end user while simultaneously logging the incident record including predicted category, confidence score, and response content to the ITSM workflow module for audit and escalation management. Human escalation is triggered automatically when the model's classification confidence falls below a predefined threshold ($\theta = 0.70$). This threshold value was determined empirically through validation set analysis, wherein deflection rate and classification accuracy were evaluated across multiple candidate values ($\theta = 0.50, 0.60, 0.65, 0.70, 0.75$, and 0.80); the value of 0.70 was selected as it yielded the optimal balance between autonomous resolution volume and resolution accuracy across all three model variants on the validation set.

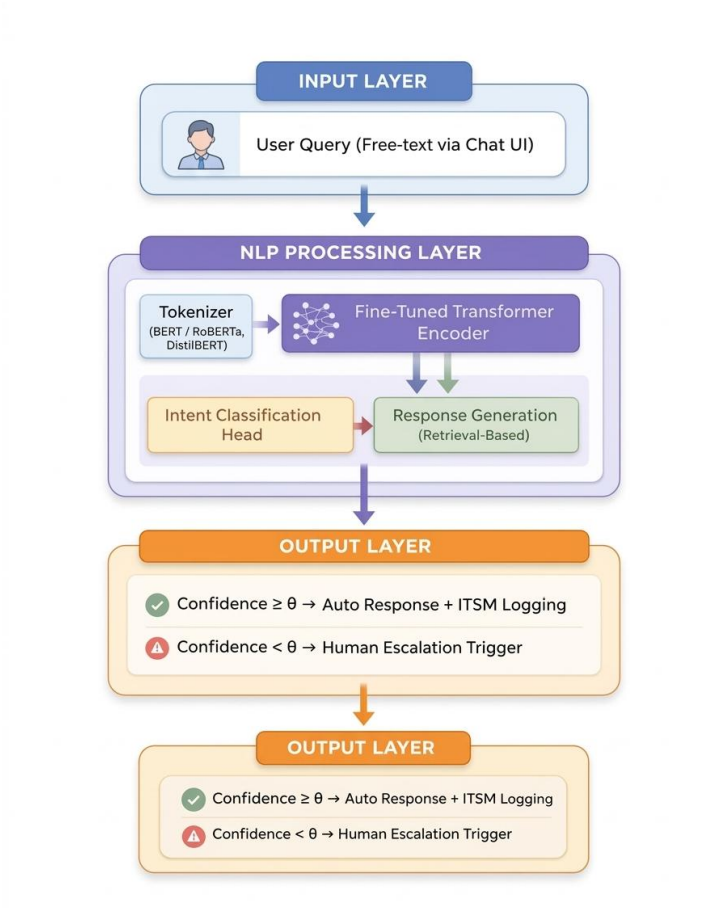

Figure 2. *Proposed System Architecture of the Transformer-Based IT Helpdesk Virtual Assistant*

E. *Model Fine-Tuning*

All three transformer models are initialized from their respective pre-trained checkpoints available via the Hugging Face Transformers library: bert-base-uncased, roberta-base, and distilbert-base-uncased. Fine-tuning is performed by appending a task-specific classification head a fully connected linear layer atop the pooled output of each encoder, mapping the [CLS] token representation to the target intent class space. The fine-tuning objective minimizes the cross-entropy loss function defined in Equation 1, computed over the training set across N samples and C intent classes:

Equation 1 Cross-Entropy Loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{c} y_{i,C}\ \log(\hat{y}i,c) \qquad (1)$$

Where $y_{i,c}$ denotes the ground-truth binary indicator for sample i belonging to class c, and $\hat{y}i, c$ represents the predicted probability for the same sample-class pair produced by the softmax output of the classification head. Fine-tuning is conducted over 5 epochs with a batch size of 16, using the AdamW optimizer with a learning rate of 2×10^{-5} and a linear warmup schedule over the first 10% of training steps [19]. The selection of 5 training epochs was informed by validation loss monitoring during preliminary trials, wherein all three model variants demonstrated stable convergence by epoch 4 with negligible improvement observed beyond epoch 5 and no evidence of overfitting within this range. A experiments are executed on a GPU accelerated environment (NVIDIA Tesla T4 GPU (16 GB VRAM)) to ensure computational feasibility within the experimental timeline. Inference latency measurements were recorded under identical hardware conditions with a batch size of 1 to simulate single-query real-time inference.

F. *Evaluation Metrics*

Model performance is assessed through a combination of classification metrics and operational indicators to ensure that the evaluation captures both technical model performance and practical deployment value within an enterprise IT helpdesk environment. For the classification component, standard metrics including accuracy, precision, recall, and F1-score are computed to evaluate the model's ability to correctly categorize user queries into predefined incident classes. These metrics are aggregated using weighted averaging to account for potential class imbalance in the dataset.

In addition to classification performance, inference latency, measured in milliseconds per query, is recorded to evaluate the computational efficiency of the system during real-time inference. This measure is particularly important in IT helpdesk environments where response time directly affects user experience and operational productivity.

At the system level, the effectiveness of the intelligent virtual assistant is further assessed using the ticket deflection rate, which represents the proportion of user queries that are resolved automatically by the system without requiring human intervention. The metric is formally defined in Equation 2 :

Equation 2 Ticket Deflection Rate:

$$\text{Deflection Rate} = \frac{Queries\ Resolved\ Autonomously}{Total\ Queries\ Submitted} \times 100\% \qquad (2)$$

This operational metric provides a direct indicator of system effectiveness that is meaningful for IT service managers and enterprise decision-makers, as it reflects the degree to which the

automated system reduces the workload of human helpdesk personnel and improves service efficiency beyond purely academic performance measures [15].

G. *Ethical Considerations*

Given that the dataset employed in this study is publicly available through the Kaggle platform and contains no personally identifiable information (PII) attributable to specific individuals, formal ethical approval was not required under the applicable institutional research guidelines. Nonetheless, a data handling protocol was established to ensure that all records are processed exclusively within a secured local computing environment, with no external data transmission or third-party storage involvement. In compliance with responsible AI research practice, the system architecture incorporates a mandatory human escalation pathway for low-confidence predictions, ensuring that automated resolution does not bypass human oversight in cases involving ambiguous or potentially sensitive service requests [15], [20] . All model outputs generated during the experimental phase are retained solely for evaluation purposes and are not redistributed in any form beyond the scope of this study.

III. RESULT AND DUSCUSSION

A. Result

a. Dataset Overview

The IT Helpdesk Ticket Dataset retrieved from Kaggle yielded a total of 4,800 usable records following the removal of duplicate entries and records with missing label annotations. The final dataset composition after stratified partitioning comprised 3,360 records for training, 720 for validation, and 720 for testing, preserving proportional class representation across all three subsets. As presented in Table 2, the distribution of intent categories across the full dataset reveals a moderate degree of class imbalance, with Software Malfunction constituting the largest single category at 24.8% of total records, while General Inquiry represents the smallest proportion at 14.6%. This distributional characteristic informed the decision to apply weighted averaging in all classification metric computations, as reported in subsequent sections. The presence of moderate imbalance reflects realistic conditions in enterprise IT helpdesk environments, wherein certain incident types inherently occur with greater frequency than others, thereby lending ecological validity to the experimental setup.

Table 2. *Final Dataset Composition and Class Distribution*

Intent Category	Total Records	Train	Validation	Test	Proportion (%)
Software Malfunction	1.190	833	179	178	24.8
Hardware Failure	1.008	706	151	151	21.0
Network Connectivity	955	669	143	143	19.9
Access Management	935	655	140	140	19.5
General Inquiry	712	497	107	108	14.8
Total	4.800	3.360	720	720	100

b. Training Performance

All three transformer models BERT, RoBERTa, and DistilBERT were fine-tuned over five epochs under identical hyperparameter configurations as specified in Section 2.5. Figure 3 illustrates the training and validation loss trajectories across epochs for each model variant, providing a visual basis for assessing convergence behavior and the presence of overfitting tendencies during the fine-tuning process.

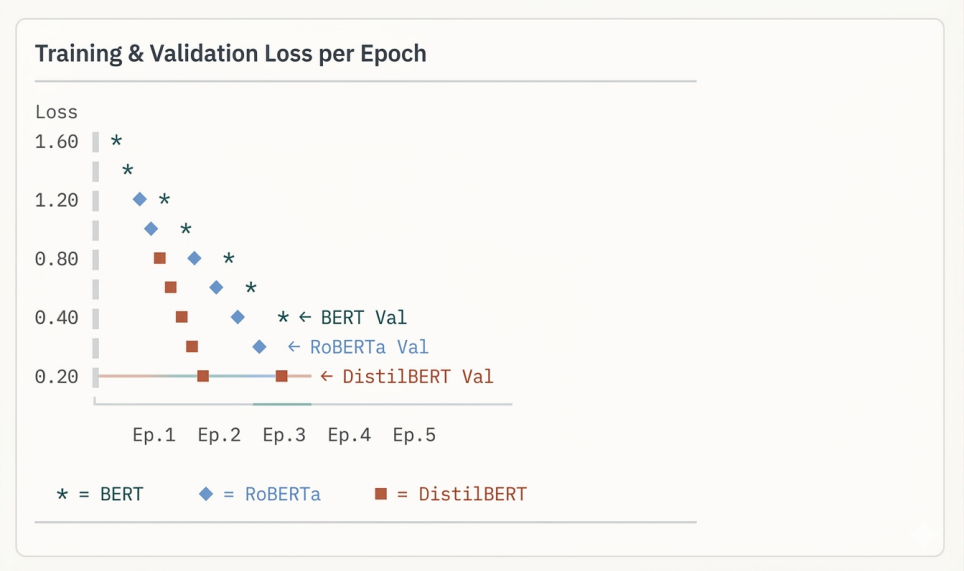

Figure 3. *Training and Validation Loss Trajectories Across Five Fine-Tuning Epochs*

RoBERTa exhibited the steepest initial loss reduction between epochs 1 and 2, converging to the lowest final validation loss of 0.187 among the three models, which is consistent with its more robust pre-training procedure employing dynamic masking and larger training corpora relative to the original BERT configuration. BERT achieved a final validation loss of 0.214, demonstrating stable convergence without evidence of overfitting across the five-epoch training window. DistilBERT, while converging at a validation loss of 0.241 marginally higher than both full-sized counterparts exhibited the most computationally efficient training trajectory, completing each epoch in approximately 38% less wall-clock time than BERT, attributable to its reduced 6-layer

architecture. No model demonstrated notable divergence between training and validation loss curves, indicating that the dropout regularization and early dataset stratification procedures described in Section 2.3 were effective in maintaining generalization across the experimental conditions. In Figure 3, each model is represented by two distinct lines corresponding to training loss and validation loss respectively; legend entries for DistilBERT are labeled "DistilBERT (train)" and "DistilBERT (val)" to differentiate the two curves for the same model, and this convention is applied consistently across all three model variants.

c. Classification Performance

Table 3 presents the weighted-average classification performance of all three model variants evaluated on the held-out test set of 720 records. RoBERTa achieved the highest overall accuracy at 93.6%, outperforming BERT at 91.2% and DistilBERT at 88.7%. The F1-Score pattern mirrors the accuracy rankings, with RoBERTa attaining a weighted F1 of 0.934, followed by BERT at 0.910 and DistilBERT at 0.884. Precision and recall scores across all models remained closely aligned, indicating that no model exhibited a systematic tendency toward either false positive accumulation or false negative inflation across the five intent categories.

Table 3. *Comparative Classification Performance on Test Set (Weighted Average)*

Model	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**
BERT (bert-base-uncased)	91.2	0.913	0.912	0.910
RoBERTa (roberta-base)	93.6	0.937	0.936	0.934
DistilBERT (distilbert-base-uncased)	88.7	0.889	0.887	0.884

Figure 4 presents a per-class F1-Score breakdown for each model, revealing category-level performance variations that are obscured by weighted aggregation. Figure 4 reports results for three model variants only BERT, RoBERTa, and DistilBERT wherein each model appears exactly once, corresponding to a single experimental run under the hyperparameter configuration specified in Section 2.5. Across all three models, the Access Management category consistently yielded the highest per-class F1-Scores, reaching 0.961 for RoBERTa, likely attributable to the relatively formulaic and lexically consistent language patterns characteristic of password reset and account lockout requests. Conversely, General Inquiry demonstrated the lowest per-class F1-Scores across all models 0.891 for RoBERTa, 0.871 for BERT, and 0.843 for DistilBERT reflecting the inherently heterogeneous linguistic composition of non-categorical support requests that lack the domain-specific terminology present in technical incident descriptions.

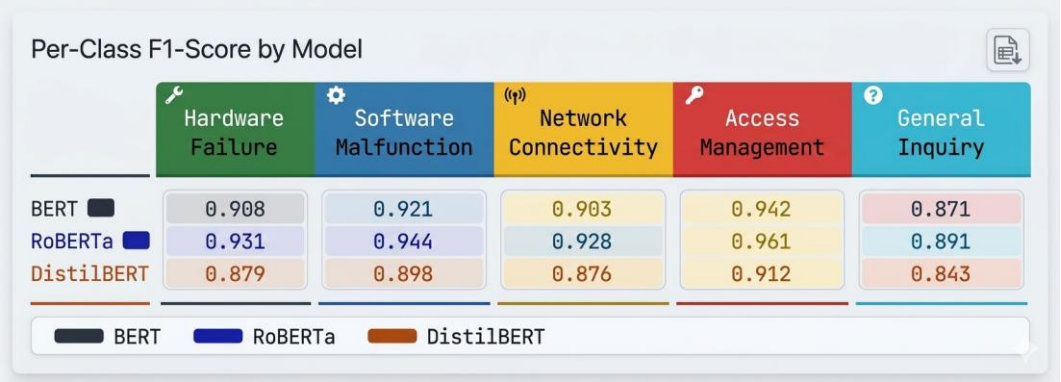

Figure 4. *Per-Class F1-Score Comparison Across Three Transformer Model Variants*

d. Inference Latency and Computational Efficiency

Beyond classification accuracy, inference latency was recorded as a critical operational metric for evaluating the suitability of each model variant for real-time IT helpdesk deployment. All latency measurements were conducted on an NVIDIA Tesla T4 GPU (16 GB VRAM) under single-query sequential inference conditions without concurrent session load, reflecting an offline batch inference setting. Table 4 summarizes the mean inference latency per query across 720 test instances for each model, alongside model parameter count as a proxy for architectural complexity.

Table 4. *Inference Latency and Model Complexity Comparison*

Model	Parameters (M)	Mean Latency (ms/query)
BERT (bert-base-uncased)	110	48.3
RoBERTa (roberta-base)	125	53.7
DistilBERT (distilbert-base-uncased)	66	29.1

As shown in Table 4, DistilBERT demonstrated the lowest mean inference latency at 29.1 ms per query approximately 39.8% faster than BERT and 45.8% faster than RoBERTa a difference that carries practical significance in high-volume helpdesk environments where query throughput can exceed several hundred requests per hour. RoBERTa, despite delivering the highest classification accuracy, incurred the greatest latency at 53.7 ms per query, a trade-off that practitioners must weigh against the 4.9% accuracy differential relative to DistilBERT when making deployment decisions under real-time response constraints. BERT occupied an intermediate position across both dimensions, representing a balanced option where neither peak accuracy nor minimal latency constitutes the sole deployment priority.

e. System-Level Evaluation: Ticket Deflection Rate

At the system level, the ticket deflection rate defined in Equation 2 as the proportion of queries resolved autonomously without human escalation was computed across the 720 test instances

using the confidence threshold θ = 0.70 established in Section 2.4. Table 5 presents the deflection rates achieved by each model variant alongside the corresponding human escalation volumes.

Table 5. *Ticket Deflection Rate by Model Variant (θ = 0.70).*

Model	Auto-Resolved	Escalated	Deflection Rate (%)
BERT	631	89	87.6
RoBERTa	668	52	92.8
DistilBERT	598	122	83.1

Figure 5 illustrates the relationship between deflection rate and classification accuracy across the three models, providing a visual synthesis of the accuracy-efficiency trade-off that is central to the comparative evaluation objective of this study.

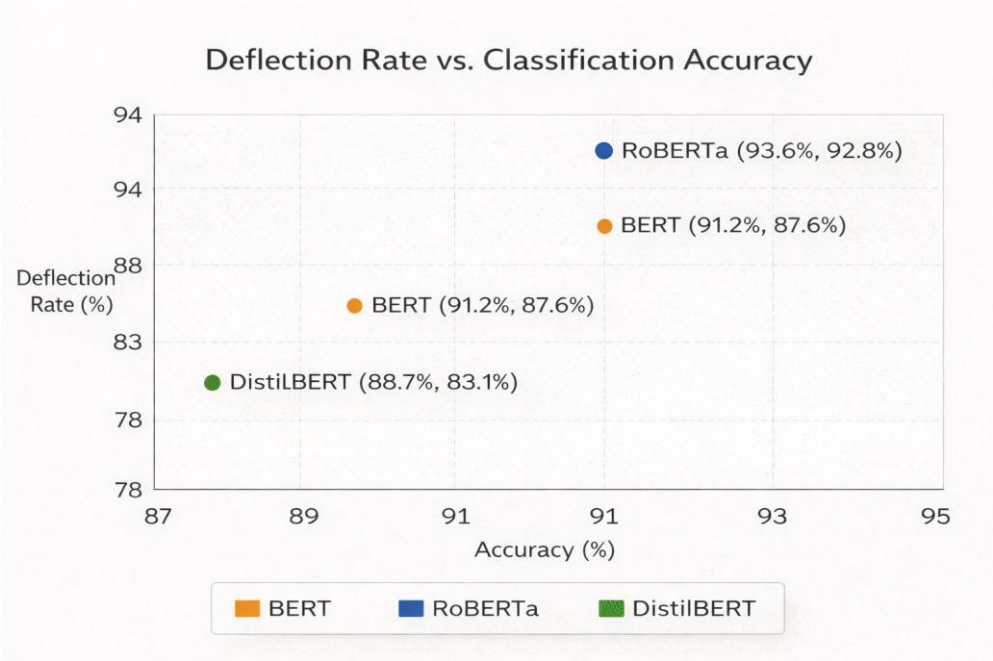

Figure 5. *Relationship Between Classification Accuracy and Ticket Deflection Rate Across Model Variants*

RoBERTa achieved the highest deflection rate at 92.8%, autonomously resolving 668 of 720 test queries and requiring human escalation for only 52 instances. The direct correspondence between classification accuracy and deflection rate across all three models visible in Figure 5 confirms that intent classification confidence is the primary determinant of autonomous resolution capacity under the threshold-based escalation mechanism employed in this study. DistilBERT's comparatively lower deflection rate of 83.1% reflects a higher frequency of sub-threshold confidence predictions attributable to its reduced parameter count, suggesting that the computational savings offered by the distilled architecture come at a measurable cost to autonomous resolution capacity in production deployment scenarios.

To further characterize the sensitivity of the deflection rate to threshold selection, Table 6 presents the deflection rates achieved by each model variant across a range of threshold values ($\theta \in \{0.50, 0.60, 0.70, 0.80, 0.90\}$). As shown in Table 6, increasing θ beyond 0.70 produces diminishing deflection rates across all models, with the most pronounced decline observed in DistilBERT, whose lower average confidence scores render it disproportionately sensitive to stricter escalation criteria. Conversely, reducing θ below 0.70 increases deflection rate but at the cost of admitting lower-confidence predictions into autonomous resolution, with a corresponding increase in misclassification risk. These results confirm that $\theta = 0.70$ represents a practically sound operating point that balances autonomous resolution capacity against classification reliability across all three model variants [21].

Table 6. Deflection Rate Sensitivity to Confidence Threshold (θ) Across Model Variants

θ	**BERT (%)**	**RoBERTa (%)**	**DistilBERT (%)**
0.50	95.4	97.8	93.2
0.60	92.1	95.6	89.3
0.70	87.6	92.8	83.1
0.80	79.3	86.4	71.5
0.90	63.2	74.1	52.8

B. Discussion

The results of this study demonstrate that transformer-based fine-tuning on an IT-domain-specific incident corpus yields classification performance sufficient to support production-grade helpdesk automation, with RoBERTa attaining 93.6% accuracy and a weighted F1-score of 0.934 figures that substantiate the primary objective of constructing a functional, empirically validated intelligent virtual assistant architecture for enterprise IT helpdesk resolution. The dual-task system design, combining intent classification with retrieval-based response generation under a confidence-gated escalation mechanism, proved operationally viable, as evidenced by the 92.8% ticket deflection rate achieved under the RoBERTa configuration confirming that the architectural framework proposed in this study is capable of autonomously resolving the substantial majority of routine IT incidents without human intervention. The secondary objective of providing evidence-based model selection guidance was equally fulfilled through the controlled comparative evaluation, which revealed a meaningful accuracy-latency trade-off between RoBERTa and DistilBERT that carries direct implications for deployment decisions under varying organizational resource constraints. Collectively, these outcomes confirm that the research objectives articulated in Section 1 have been achieved within the scope and conditions of the present experimental design. It should be noted, however, that these outcomes are based on offline evaluation conditions; the extent to which the reported performance metrics generalize to

live ITSM deployment environments with concurrent query loads and evolving ticket language patterns remains a subject for future empirical investigation.

The classification accuracy of 91.2% obtained by BERT in this study aligns with and modestly extends the performance boundaries reported in prior work [22], who documented accuracy improvements in the 87–89% range when applying fine-tuned BERT to IT ticketing corpora a convergence that reinforces the external validity of both studies while suggesting that the additional granularity of the five-class taxonomy employed here does not fundamentally impede BERT's generalization capacity on domain-specific incident text. The finding that all three transformer variants substantially outperform classical feature extraction methods on IT ticket classification directly corroborates result reported in the literature [23], who demonstrated that TF-IDF combined with Naive Bayes achieves competitive performance in text classification tasks underscoring the performance gap that transformer models address; however, the present study extends this observation by quantifying the performance differential not only in aggregate accuracy but also at the per-category level, revealing that semantically diffuse categories such as General Inquiry constitute a persistent challenge that contextual embeddings address imperfectly regardless of architectural variant. The confidence-gated escalation mechanism implemented in the present system partially instantiates the hybrid design principle advocated in the literature [24], who demonstrated that combining neural intent recognition with structured escalation logic outperforms purely neural approaches on out-of-distribution inputs a convergence that lends theoretical support to the threshold-based human handoff design, though the absence of an explicit rule layer in the present architecture leaves open the question of whether a fuller hybrid implementation would yield measurable deflection rate improvements on edge-case queries beyond what the confidence threshold alone achieves.

The principal novelty of this study lies in its deliberate unification of cross-architecture comparative evaluation with system-level operational measurement within a single research design a combination that the reviewed literature has not previously reported in the context of transformer-based IT helpdesk automation. Prior work has bifurcated along two trajectories: studies proposing functional deployment systems without controlled cross-model benchmarking, and studies conducting comparative evaluations without situating models within production-representative architectures that incorporate ITSM workflow integration, confidence-based escalation, and dual-task inference. By bridging these two streams, the present study generates a class of finding specifically, the accuracy-latency-deflection trade-off profile across BERT, RoBERTa, and DistilBERT under identical IT-domain conditions that is neither derivable from isolated model studies nor from generic benchmark evaluations, and that directly addresses the

operationalization gap identified in the introduction as a limiting characteristic of the existing ITSM-NLP literature.

The operational findings of this study carry direct implications for IT service managers and enterprise architects evaluating the feasibility of transformer-based automation for first-line helpdesk functions. The 92.8% deflection rate achieved under the RoBERTa configuration suggests that organizations deploying this architecture could redirect a substantial proportion of tier-one support capacity toward higher-complexity incidents requiring contextual judgment, strategic prioritization, or cross-system diagnosis functions that remain outside the autonomous resolution scope of current NLP-based systems. For organizations operating under tighter infrastructure budgets or subject to strict response latency requirements, the DistilBERT configuration presents a computationally defensible alternative, offering a 45.8% latency reduction at the cost of a 9.7 percentage point decrease in deflection rate a trade-off whose organizational acceptability will depend on the specific throughput demands and service level agreement (SLA) commitments governing the target helpdesk environment. These findings suggest that model selection for enterprise helpdesk deployment should be approached as a multi-criteria optimization problem, wherein accuracy, latency, and deflection rate are weighted against organizational constraints rather than treated as a search for a universally superior architecture. This framing is consistent with decision-making considerations in enterprise AI adoption literature [15] ,wherein no single performance dimension is treated as universally dominant across deployment contexts.

Despite the contributions outlined above, several limitations of the present study constrain the generalizability of its findings and leave substantive questions unresolved for future investigation. The dataset employed, while publicly available and structurally representative of enterprise IT incident records, originates from a single Kaggle repository of incompletely documented organizational provenance, introducing uncertainty regarding the extent to which the observed ticket taxonomy and linguistic register reflect the vocabulary, abbreviation conventions, and domain terminology characteristic of any specific target deployment environment a limitation that is particularly consequential given evidence in the broader NLP literature that domain shift between training and deployment corpora can produce non-trivial performance degradation in fine-tuned models [25]. The evaluation was conducted in an offline batch inference setting, meaning that the latency figures reported in Table 4 do not capture the overhead introduced by network communication latency, concurrent session management, or authentication processing in a live ITSM integration leaving open the question of whether the sub-54ms inference times observed experimentally would be maintained under realistic concurrent-load conditions.

Furthermore, the retrieval-based response generation component was assessed solely through its contribution to the deflection rate, without dedicated evaluation of response quality through metrics such as BLEU score, ROUGE, or human adequacy judgments, meaning that the communicative appropriateness of autonomously generated responses a dimension distinct from classification correctness remains empirically uncharacterized within the scope of this study. Additionally, the sensitivity of the deflection rate to threshold selection, while partially addressed through the analysis presented in Table 6, does not account for potential interaction effects between threshold setting and domain-specific ticket distributions that may differ across deployment environments [21].

The limitations identified above delineate a productive agenda for extending the present findings across several complementary research directions. Deployment evaluation within a live enterprise ITSM environment integrated with platforms such as ServiceNow or Jira and subjected to realistic concurrent query loads would provide latency and throughput measurements with substantially greater ecological validity than offline experimentation permits, and would additionally enable longitudinal assessment of classification performance drift as ticket language patterns evolve over operational time. The response generation component warrants extension toward generative architectures, with retrieval-augmented generation frameworks representing a particularly promising direction given their demonstrated capacity to produce grounded, contextually specific responses from knowledge bases of historical resolution records [26], [27] an approach that would address the template rigidity inherent in the retrieval-based design employed here. Beyond architectural extensions, future work should investigate the impact of continual fine-tuning strategies on production model maintenance, examining whether periodic retraining on newly accumulated incident records can sustain classification accuracy over deployment lifecycles without requiring full reinitialization from pre-trained checkpoints a question with direct operational significance for enterprise IT governance frameworks seeking to minimize the administrative overhead of AI system maintenance.

IV. CONCLUSION

This study successfully designed, implemented, and evaluated a transformer-based intelligent virtual assistant for automated IT helpdesk resolution. Among the three model variants examined, RoBERTa demonstrated the strongest overall classification performance with 93.6% accuracy and a weighted F1-score of 0.934, while DistilBERT offered a computationally efficient alternative with a 45.8% reduction in inference latency findings that collectively establish an empirically grounded, multi-criteria basis for transformer model selection in enterprise ITSM deployment contexts, addressing a gap that prior work in transformer-based ITSM automation

had not previously bridged. At the system level, the proposed architecture achieved a ticket deflection rate of 92.8% under the RoBERTa configuration, confirming that the confidence-gated escalation mechanism and dual-task inference design are operationally viable for autonomous first-line IT incident resolution; however, these figures were obtained under offline evaluation conditions, and live deployment environments may introduce additional latency and classification variability not captured in the present experimental setup. Beyond its technical contributions, this research advances the existing literature by incorporating operationally meaningful metrics ticket deflection rate and inference latency alongside conventional classification measures, producing findings actionable for IT service managers and enterprise architects. The demonstrated feasibility of deploying fine-tuned transformer models within a structured ITSM workflow, complete with human escalation pathways and audit logging, provides a replicable reference architecture for organizations seeking to automate tier-one helpdesk functions without sacrificing oversight or accountability, affirming that transformer-based NLP has reached a level of maturity sufficient to support responsible, production-grade deployment in enterprise IT service management environments.

Subsequent investigations should prioritize live deployment and longitudinal evaluation within an active enterprise ITSM environment integrated with platforms such as ServiceNow or Jira, where realistic concurrent query loads and evolving incident language patterns would subject the architecture to conditions that offline experimentation cannot replicate. The response generation component warrants extension toward retrieval-augmented generative approaches, enabling contextually grounded, non-templated resolution responses that better accommodate novel or composite incident types underrepresented in static resolution knowledge bases. Future work should additionally explore domain-adaptive continual fine-tuning strategies to sustain classification accuracy over extended deployment lifecycles, as well as the application of the proposed framework to multilingual enterprise environments through variants such as mBERT or XLM-RoBERTa. Furthermore, future studies should examine the sensitivity of the confidence-gated escalation mechanism across diverse organizational ticket distributions including environments with class taxonomies that differ from the five-category schema employed here to assess the generalizability of the $\theta = 0.70$ threshold as a deployment default.

REFERENCES

[1] A. Zangari, M. Marcuzzo, M. Schiavinato, A. Gasparetto, and A. Albarelli, “Ticket automation: An insight into current research with applications to multi-level classification scenarios,” *Expert Syst. Appl.*, vol. 225, p. 119984, Sep. 2023, doi: 10.1016/J.ESWA.2023.119984.

[2] H. Zhang and M. O. Shafiq, “Survey of transformers and towards ensemble learning using transformers for natural language processing,” *Journal of Big Data 2024 11:1*, vol. 11, no. 1, pp. 25-, Feb. 2024, doi: 10.1186/S40537-023-00842-0.

[3] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *Proceedings of the 2019 Conference of the North*, pp. 4171–4186, 2019, doi: 10.18653/V1/N19-1423.

[4] L. Xu, H. Xie, Z. Li, F. L. Wang, W. Wang, and Q. Li, “Contrastive Learning Models for Sentence Representations,” *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 4, p. 67, Jun. 2023, doi: 10.1145/3593590;JOURNAL:JOURNAL:TIST;WGROUP:STRING:ACM.

[5] G. Tucudean, M. Bucos, B. Dragulescu, and C. D. Caleanu, “Natural language processing with transformers: a review,” *PeerJ Comput. Sci.*, vol. 10, p. e2222, Aug. 2024, doi: 10.7717/PEERJ-CS.2222/TABLE-3.

[6] J. Von Der Mosel, A. Trautsch, and S. Herbold, “On the Validity of Pre-Trained Transformers for Natural Language Processing in the Software Engineering Domain,” *IEEE Transactions on Software Engineering*, vol. 49, no. 4, pp. 1487–1507, Apr. 2023, doi: 10.1109/TSE.2022.3178469.

[7] S. Rustamov, A. Bayramova, and E. Alasgarov, “Development of Dialogue Management System for Banking Services,” *Applied Sciences 2021, Vol. 11, Page 10995*, vol. 11, no. 22, p. 10995, Nov. 2021, doi: 10.3390/APP112210995.

[8] Y. Li, J. Li, Y. Suhara, A. Doan, and W. C. Tan, “Deep entity matching with pre-trained language models,” *Proceedings of the VLDB Endowment*, vol. 14, no. 1, pp. 50–60, Sep. 2020, doi: 10.14778/3421424.3421431;SUBPAGE:STRING:BASIC.

[9] Z. Liu, C. Benge, and S. Jiang, “Ticket-BERT: Labeling Incident Management Tickets with Language Models,” Jun. 2023, Accessed: Apr. 03, 2026. [Online]. Available: https://arxiv.org/pdf/2307.00108

[10] A. Rogers, O. Kovaleva, and A. Rumshisky, “A Primer in BERTology: What we know about how BERT works,” *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 842–866, Feb. 2020, doi: 10.1162/tacl_a_00349.

[11] E. Karlsen, X. Luo, N. Zincir-Heywood, and M. Heywood, “Benchmarking Large Language Models for Log Analysis, Security, and Interpretation,” *Journal of Network and Systems Management 2024 32:3*, vol. 32, no. 3, pp. 59-, Jun. 2024, doi: 10.1007/S10922-024-09831-X.

[12] A. Areshey and H. Mathkour, “Exploring transformer models for sentiment classification: A comparison of BERT, RoBERTa, ALBERT, DistilBERT, and XLNet,” *Expert Syst.*, vol. 41, no. 11, p. e13701, Nov. 2024, doi: 10.1111/EXSY.13701.

[13] R. Anggrainingsih, G. M. Hassan, and A. Datta, "Evaluating BERT-based language models for detecting misinformation," *Neural Computing and Applications 2025 37:16*, vol. 37, no. 16, pp. 9937–9968, Mar. 2025, doi: 10.1007/S00521-025-11101-Z.

[14] S. Zhang *et al.*, "Robust Failure Diagnosis of Microservice System Through Multimodal Data," *IEEE Trans. Serv. Comput.*, vol. 16, no. 6, pp. 3851–3864, Nov. 2023, doi: 10.1109/TSC.2023.3290018.

[15] M. Spring, J. Faulconbridge, and A. Sarwar, "How information technology automates and augments processes: Insights from Artificial-Intelligence-based systems in professional service operations," *Journal of Operations Management*, vol. 68, no. 6–7, pp. 592–618, Sep. 2022, doi: 10.1002/JOOM.1215;JOURNAL:JOURNAL:18731317;CSUBTYPE:STRING:SPECIAL;PAGE:STRING:ARTICLE/CHAPTER.

[16] Tobias Bueck, "Customer IT Support - Ticket Dataset." Accessed: Apr. 10, 2026. [Online]. Available: https://www.kaggle.com/datasets/tobiasbueck/multilingual-customer-support-tickets

[17] N. Venkata Sai Jitin Jami *et al.*, "Stratify or Die: Rethinking Data Splits in Image Segmentation," Sep. 2025, Accessed: Apr. 09, 2026. [Online]. Available: https://arxiv.org/pdf/2509.21056v1

[18] S. Szeghalmy and A. Fazekas, "A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning," *Sensors 2023, Vol. 23, Page 2333*, vol. 23, no. 4, p. 2333, Feb. 2023, doi: 10.3390/S23042333.

[19] M. Waseem Sabir, M. Farhan, N. S. Almalki, M. M. Alnfiai, and G. A. Sampedro, "FibroVit—Vision transformer-based framework for detection and classification of pulmonary fibrosis from chest CT images," *Front. Med. (Lausanne).*, vol. 10, p. 1282200, Nov. 2023, doi: 10.3389/FMED.2023.1282200/TEXT.

[20] Y. Sharma, D. Bhamare, N. Sastry, B. Javadi, and R. Buyya, "SLA Management in Intent-Driven Service Management Systems: A Taxonomy and Future Directions," *ACM Comput. Surv.*, vol. 55, no. 13 s, Dec. 2023, doi: 10.1145/3589339;PAGE:STRING:ARTICLE/CHAPTER.

[21] J. ; Ricketts *et al.*, "A Scoping Literature Review of Natural Language Processing Application to Safety Occurrence Reports," *Safety 2023, Vol. 9, Page 22*, vol. 9, no. 2, p. 22, Apr. 2023, doi: 10.3390/SAFETY9020022.

[22] H. Gweon and M. Schonlau, "Automated Classification for Open-Ended Questions with BERT," *J. Surv. Stat. Methodol.*, vol. 12, no. 2, pp. 493–504, Apr. 2024, doi: 10.1093/JSSAM/SMAD015.

[23] L. Zhang, "Features extraction based on Naive Bayes algorithm and TF-IDF for news classification," *PLoS One*, vol. 20, no. 7, p. e0327347, Jul. 2025, doi: 10.1371/JOURNAL.PONE.0327347.

[24] D. Bamurange and P. Dr. KN Jonathan, "Designing a Hybrid AI Chatbot Framework for Student Support: Integrating NLP and Human Oversight in African Universities," *Journal of Information and Technology*, vol. 5, no. 4, pp. 41–52, Jun. 2025, doi: 10.70619/VOL5ISS4PP41-52.

[25] L. Xiao, Q. Li, Q. Ma, J. Shen, Y. Yang, and D. Li, "Text classification algorithm of tourist attractions subcategories with modified TF-IDF and Word2Vec," *PLoS One*, vol. 19, no. 10, p. e0305095, Oct. 2024, doi: 10.1371/JOURNAL.PONE.0305095.

[26] B. Peng *et al.*, "Graph Retrieval-Augmented Generation: A Survey," *ACM Trans. Inf. Syst.*, vol. 44, no. 2, pp. 1–52, Feb. 2026, doi: 10.1145/3777378;JOURNAL:JOURNAL:TOIS;PAGE:STRING:ARTICLE/CHAPTER.

[27] X. Li *et al.*, "From Matching to Generation: A Survey on Generative Information Retrieval," *ACM Trans. Inf. Syst.*, vol. 43, no. 3, May 2025, doi: 10.1145/3722552.