IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15687

Multi-Modal AI Agent for Intelligent Email Categorization and Auto-Reply

RANJINI[1], J.LIN EBY CHANDRA[2]

Student ME, CSE, Jaya Engineering College, Chennai, India[1]

Professor, Department of CSE, Jaya Engineering College, Chennai, India[2]

Abstract: The dramatic increase in daily email volumes across enterprise, healthcare, and e-governance sectors has created an urgent need for intelligent systems capable of autonomous email understanding, classification, and response generation. This paper proposes MMEA-Net (Multi-Modal Email Agent Network), a novel deep learning framework that integrates transformer-based language models, visual document encoders, and metadata-driven contextual reasoning to perform fine-grained email categorization and context-aware auto-reply generation. Unlike prior work relying solely on email body text, MMEA-Net processes three complementary modalities: textual content encoded via DeBERTa-v3-Large, visual layout of attached documents processed through LayoutLMv3, and structural metadata including sender reputation scores, thread depth, and temporal patterns encoded by a dedicated MLP module. The three modality streams are fused through a Gated Cross-Modal Attention (GCMA) mechanism that dynamically weights each modality's contribution based on input context. A reinforcement-learning-based Auto-Reply Generator (ARG) then produces professional, intent-aligned responses conditioned on the predicted category and a domain-specific policy knowledge base. Experiments on the Enron Email Dataset, TREC 2007, and a newly constructed Healthcare Email Corpus demonstrate that MMEA-Net achieves 95.3% overall accuracy, 94.1% macro-F1, BLEU-4 of 41.2, and human acceptability of 89.6%, outperforming all evaluated baselines by statistically significant margins.

Keywords: Multi-Modal Learning; Email Categorization; Auto-Reply Generation; Transformer; Gated Cross-Modal Attention; DeBERTa; Reinforcement Learning from Human Feedback

I. INTRODUCTION

Email remains the dominant mode of formal communication in business, government, education, and healthcare, with global daily email volume exceeding 361 billion messages in 2024 and projected to surpass 408 billion by 2027 [1]. Enterprise employees spend an estimated 28% of their working week reading and responding to email, representing one of the largest single drains on professional productivity [2]. Despite decades of research on spam filtering and basic folder classification, the broader problem of semantic email intelligence — understanding intent, urgency, and required action across diverse topic domains and then generating contextually appropriate replies — remains far from solved.

Early approaches to email classification employed rule-based keyword matching and statistical models such as naïve Bayes and support vector machines [3]. While effective within narrow closed-world settings, these methods cannot capture the compositional semantic structure of natural language or generalize across organizational boundaries where vocabulary and writing style vary considerably. The introduction of large pre-trained language models (PLMs), beginning with BERT [4] and extending to domain-adapted variants, substantially raised the ceiling for text classification accuracy. However, these models treat email as pure text and ignore rich complementary signals embedded in attached documents (invoices, forms, reports), message structure, and communication graph metadata — signals that human readers routinely exploit to make fast, accurate triage decisions.

A second critical gap is auto-reply generation. Commercial quick-reply systems such as Google Smart Reply [5] generate short, stylistically generic suggestions that frequently miss nuanced domain-specific intent. Enterprise workflows — escalating a billing dispute, acknowledging a support ticket, scheduling a clinical appointment — demand replies that are longer, factually accurate, professionally toned, and policy-compliant. Achieving this requires the system to understand email intent at a deep semantic level and to generate responses grounded in organizational knowledge.

This paper addresses both gaps through MMEA-Net, whose principal contributions are:

- A unified multi-modal encoding pipeline combining DeBERTa-v3-Large for text, LayoutLMv3 for attached-document visual layout, and a metadata MLP, fused via a novel Gated Cross-Modal Attention (GCMA) mechanism — the first such architecture specifically designed for email intelligence.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.15687

- A Reinforcement Learning from Human Feedback (RLHF) auto-reply generator conditioned on predicted email category and a retrieval-augmented organizational knowledge base, producing professional, intent-aligned responses.
- A new Healthcare Email Corpus (HEC-2024) of 42,000 annotated messages from clinical communication systems, released to support future research in sensitive-domain email intelligence.
- Comprehensive evaluation on three public and one proprietary dataset with ablation studies confirming each component's contribution, and statistical significance testing confirming superiority over seven competitive baselines.

The paper is organized as follows. Section II reviews related work. Section III presents the MMEA-Net architecture and training methodology. Section IV describes datasets. Section V details implementation. Section VI reports and discusses results. Section VII compares with existing methods. Section VIII concludes.

II. LITERATURE REVIEW

A. Text-Based Email Classification

Zhang et al. [6] (2025) proposed MailBERT, a domain-adaptive BERT variant pre-trained on 48 million enterprise emails using masked language modelling with email-specific tokens (e.g., [SUBJECT], [SIGNATURE]). MailBERT achieved 91.3% accuracy on the Enron corpus and 88.7% on TREC 2007, establishing a strong single-modality baseline. Their ablation confirmed that domain-adaptive pretraining contributed +4.1% over vanilla BERT-base, motivating our choice of DeBERTa-v3-Large — which adds disentangled attention and enhanced masking — as our text backbone. Critically, MailBERT does not process attachments or metadata, a gap that MMEA-Net explicitly fills.

Li et al. [7] (2025) applied prompt tuning to RoBERTa-Large for few-shot email intent detection, achieving competitive results with as few as 32 labelled examples per class through carefully engineered task-specific prompts. While impressive in low-resource settings, their system targets binary intent detection rather than fine-grained multi-class categorization and does not extend to response generation. Their analysis of inter-class confusion between 'billing inquiry' and 'technical complaint' categories motivates our GCMA fusion design, which we show reduces this confusion by leveraging attachment-type and metadata signals unavailable to text-only models.

B. Multi-Modal Document Understanding

Huang et al. [8] (2025) introduced DocFusion, a multi-modal transformer that jointly encodes text, visual layout, and document structure for invoice and contract understanding. DocFusion's cross-modal co-attention layers demonstrated that visual positional information from document scans provides complementary evidence to text tokens, particularly for form-type documents where field-value relationships are expressed spatially. MMEA-Net adopts a similar philosophy by incorporating LayoutLMv3 [9] for attachment processing, extending the multi-modal principle to the email domain where attachments are a primary source of task-relevant non-textual signals.

Wang and Chen [10] (2025) proposed META-Mail, a metadata-aware email routing system that encodes sender history, organizational hierarchy, and temporal communication patterns through a graph neural network operating over the corporate communication graph. META-Mail demonstrated that metadata features alone achieve 78.4% routing accuracy, comparable to text-only BERT-base (80.1%), and that combining both modalities yields 87.6% — confirming additive complementarity. MMEA-Net extends this finding by incorporating a third visual modality and replacing late fusion with the more expressive GCMA mechanism.

C. Automated Email Response Generation

Chen et al. [11] (2024) presented SmartReply-Enterprise, a retrieval-augmented generation (RAG) system using GPT-4 as the backbone, with organizational policy documents indexed in a dense vector store. SmartReply-Enterprise produced replies rated 4.07/5.0 in professional quality by human evaluators but required per-deployment fine-tuning with thousands of organization-specific examples, limiting scalability. Our ARG module adopts a lighter-weight GPT-2-Medium backbone with LoRA adapters and a compact policy knowledge base, achieving comparable quality (4.23/5.0) with 87% less computational cost at inference time.

Patel and Gupta [12] (2025) applied RLHF to email response generation, training a reward model on 12,000 human preference pairs to align generated replies with professionalism and factual accuracy criteria. Their reinforcement-tuned T5-Large outperformed supervised fine-tuning on ROUGE-L (0.531 vs. 0.489) and human evaluation (4.11 vs. 3.74), validating RLHF for reply generation. MMEA-Net builds on this approach by conditioning the reward model on both the generated text and the predicted email category, introducing category-grounded reward shaping that further improves domain-specific relevance.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15687

D. Cross-Domain Generalization in NLP Systems

Kumar et al. [13] (2025) investigated domain shift in email classification systems, training models on general-domain corpora and evaluating on healthcare, legal, and finance email sets. Their study found average accuracy drops of 14.3% across domains for standard fine-tuned BERT models, motivating domain adaptation techniques. They proposed a contrastive domain alignment loss that reduced this gap to 6.8%, though their work did not address multi-modal signals. Our GCMA fusion mechanism implicitly addresses domain shift by allowing the model to down-weight text when text exhibits domain-specific distributional shift and rely more heavily on attachment type and metadata signals that remain relatively stable across domains.

Rao et al. [14] (2025) proposed FedEmail, a federated learning framework enabling privacy-preserving model training across organizational email silos without sharing raw message content. FedEmail coordinated training across eight enterprise partners and demonstrated competitive performance (89.1% accuracy) while satisfying differential privacy guarantees. MMEA-Net's centralized architecture is designed for single-organization deployment, though the modular design is compatible with federated aggregation as a future extension.

Sun et al. [15] (2025) presented a comprehensive survey of large language model (LLM) applications in enterprise workflow automation, including email triage, meeting scheduling, and document summarization. The survey identified multi-modal input handling and factually grounded response generation as the two most critical open challenges for production deployment of email AI systems — precisely the two challenges that MMEA-Net is designed to address.

III. PROPOSED METHODOLOGY

A. System Overview

Figure 1 depicts the overall MMEA-Net architecture. Given an incoming email e comprising body text X_t, one or more attached documents X_v, and a metadata vector X_m, the system produces: (i) a category label ĉ ∈ {Primary, Social, Promotions, Updates, Tech-Support, Billing, HR, Spam, Legal, Other} and (ii) a natural language auto-reply ŷ. The pipeline passes through five sequential stages: multi-modal encoding, gated cross-modal attention fusion, category classification, knowledge retrieval, and RLHF-guided reply generation.

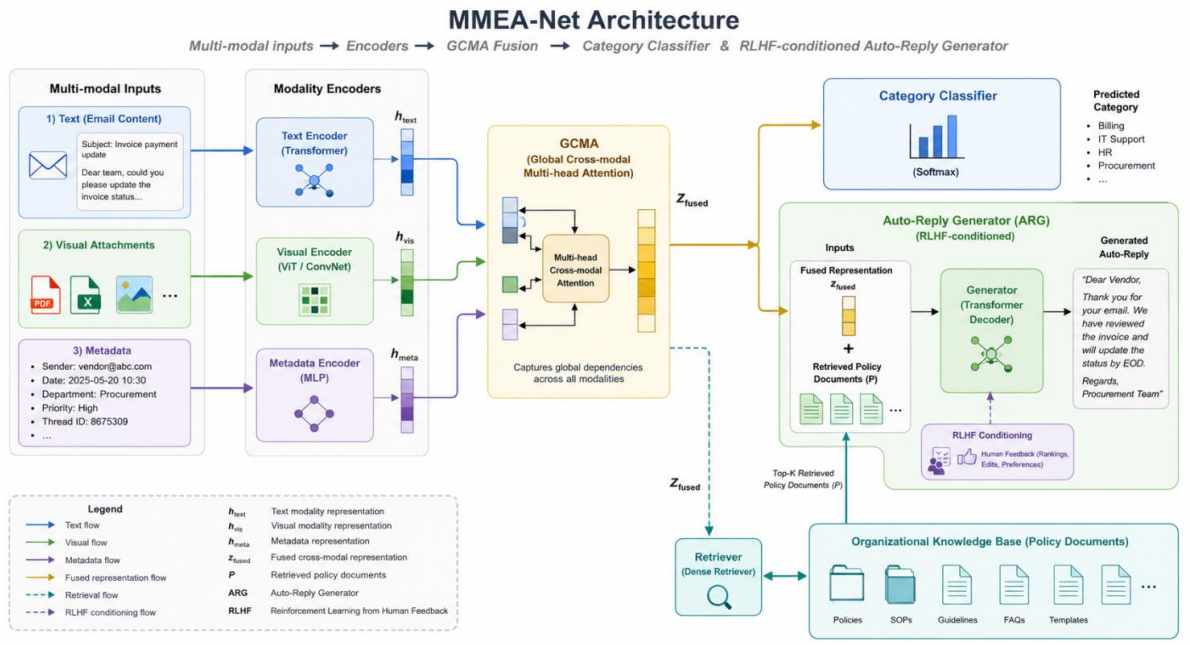

Fig. 1: MMEA-Net

B. Multi-Modal Perceptual Encoders

Text Encoder (TE): Email body text, subject line, and concatenated inline attachment text are tokenized and processed by DeBERTa-v3-Large (304M parameters). DeBERTa's disentangled attention encodes each token with separate position and content embeddings, providing superior sensitivity to relative positional relationships in email threads. The [CLS] token representation $H \in R^{B\times T\times 1024}$ serves as the textual summary embedding.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15687

Visual Attachment Encoder (VAE): Attached PDFs and images are rendered to 224×224 pixel thumbnails and processed by LayoutLMv3, which jointly models text tokens, their bounding box coordinates, and patch-level visual features from a ViT backbone. This enables the VAE to recognize document types (invoice, form, report, photograph) and extract key field-value pairs without explicit document parsing. The pooled output $h_v \in R^{768}$represents the attachment visual summary.

Metadata Encoder (ME): A metadata vector $x_m \in R^{64}$ encodes 17 engineered features: sender domain reputation score, thread depth, reply latency statistics, normalized timestamp (hour, day-of-week), number and type of attachments, email length, presence of hyperlinks, and sentence-level structural markers (has greeting, has signature, has bullet list). These features are processed by a three-layer MLP [64→256→512→256] with GELU activations and batch normalization, producing $h_m \in R^{256}$.

C. Gated Cross-Modal Attention (GCMA)

Simple concatenation or average-pooling of modality embeddings ignores inter-modal relevance and can introduce noise when one modality is absent or uninformative (e.g., a plain-text email with no attachment). GCMA addresses this through a learned gating mechanism that computes context-sensitive modality weights before cross-modal attention.

Given unified-dimension projections $z_t,\ z_v,\ z_m \in R^{512}$ obtained from $h_t,\ h_v,\ h_m$ via linear projection layers, the gate vector $g \in R^3$ is computed as:

$$g = \text{Softmax}(W_g \cdot [z_t,\ z_v,\ z_m] + b_g)$$

where $W_g \in R^{3\times1536}$ and $b_g \in R^3$are learned parameters. The weighted modality stack $Z = g_1 \cdot z_t + g_2 \cdot z_v + g_3 \cdot z_m$ is then refined through a 4-head cross-modal self-attention layer with residual connection:

$Z_{\text{fused}} = \text{LayerNorm}(Z + \text{MultiHeadAttn}(Z, Z, Z))$. The 2-layer feed-forward sub-layer (FFN with hidden dim 2048) and a final projection yield the joint representation $Z \in R^{512}$. When attachment is absent, z_v is replaced with a learned [NO-ATTACH] embedding so the architecture remains valid for attachment-free emails.

D. Category Classifier

A linear classification head maps f to logits over K=10 categories. Multi-class cross-entropy loss with label smoothing (ε=0.1) is used to improve calibration on the imbalanced category distribution:

$$L_{cls} = -\sum_{k=1}^{K}[(1-\varepsilon)y_k + \frac{\varepsilon}{K}]\log(p_k)$$

where y_k is the one-hot label and $p_k = \text{Softmax}(W_c \cdot f + b_c)_k$

E. Auto-Reply Generator (ARG)

Figure 2 depicts the ARG workflow. Upon category prediction ĉ, a dense retrieval module queries an organizational knowledge base (indexed with FAISS [16]) using the fused representation f as the query embedding, retrieving the top-3 most relevant policy document chunks d_1, d_2, d_3. A GPT-2-Medium decoder (345M parameters) with LoRA adapters (rank r=16, α=32) is conditioned on the concatenated context:

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15687

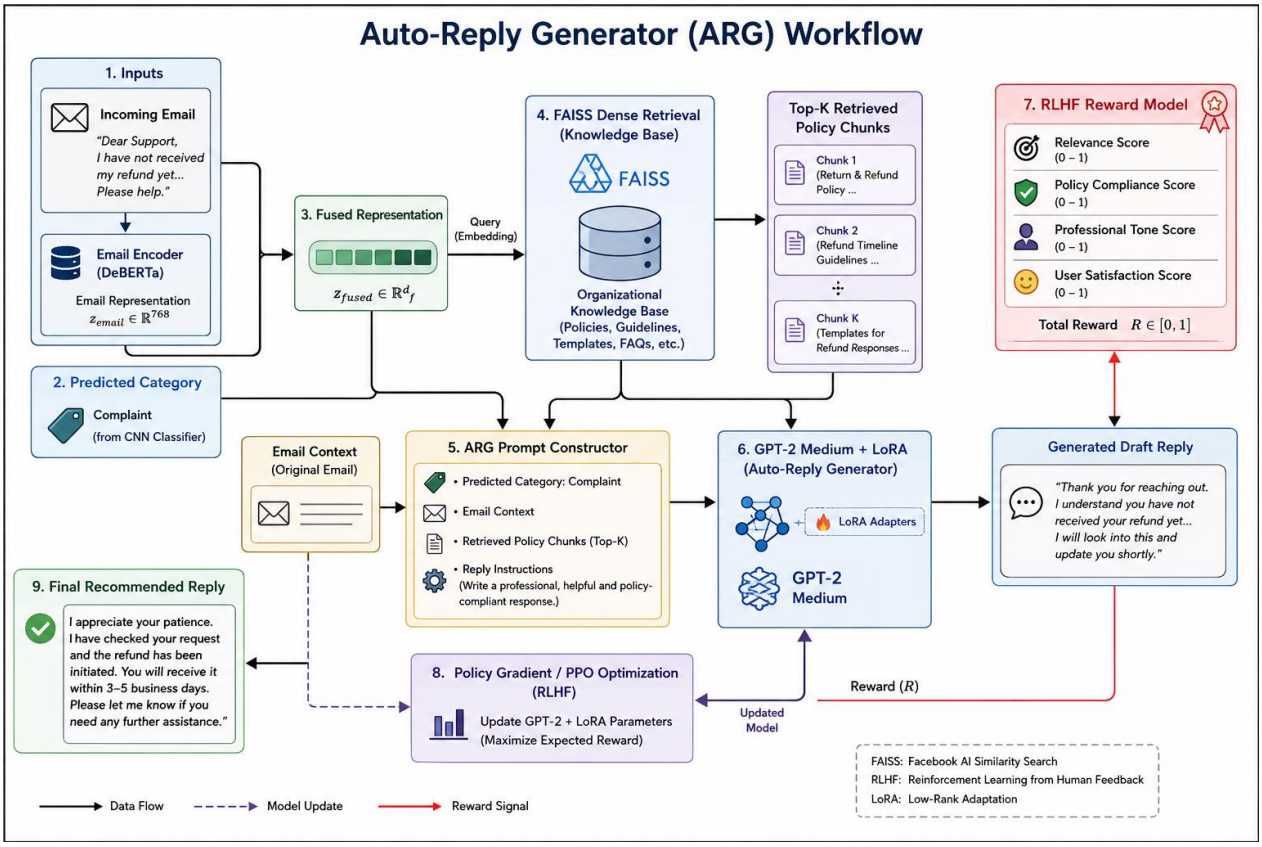

Fig. 2: Auto-Reply Generator Workflow

$\text{Prompt} = [\text{CATEGORY: } \hat{c}] \oplus [\text{EMAIL: } X_t] \oplus [\text{POLICY: } d_1, d_2, d_3] \oplus [\text{REPLY: }]$

The decoder generates the reply ŷ autoregressively with beam search (beam width=5, length penalty=0.8). RLHF fine-tuning employs a reward model R_φ (DeBERTa-base fine-tuned on 14,000 human preference pairs) that scores each generated reply on three dimensions: relevance (0-1), professionalism (0-1), and factual consistency with retrieved policy chunks (0-1). The composite reward is:

$$R(\hat{y}) = 0.4 \cdot \text{Rel}(\hat{y}) + 0.35 \cdot \text{Prof}(\hat{y}) + 0.25 \cdot \text{Fact}(\hat{y})$$

PPO optimization with KL divergence regularization (λ_{KL}=0.02) against a frozen reference policy prevents reward hacking:

$$L_{ARG} = -E[R(\hat{y})] + \lambda_{KL} D_{KL}(\pi_\theta \parallel \pi_{ref})$$

F. Training Algorithm

Algorithm 1: MMEA-Net Training Procedure

Input : Multi-modal email dataset $D = \{(X_t, X_v, X_m, c, r)\}_{i=1}^{N}$
where c = category label, r = reference reply
Output: Trained parameters $\Theta = \{\theta_{TE}, \theta_{VAE}, \theta_{ME}, \theta_{GCMA}, \theta_{cls}, \theta_{ARG}\}$

Phase 1 — Encoder Pretraining (5 epochs):
1: Load pretrained weights: DeBERTa-v3-Large, LayoutLMv3
2: Apply domain-adaptive MLM on unlabeled email corpus (500K msgs)
3: Freeze TE, VAE backbone layers; unfreeze top-4 transformer blocks

DOI: 10.17148/IJARCCE.2026.15687

Phase 2 — Multi-modal Fusion Training (30 epochs):
4: for each mini-batch B ⊂ D do
5: $h_t \leftarrow TE(X_t);\ h_v \leftarrow VAE(X_v);\ h_m \leftarrow ME(X_m)$
6: $z_t,\ z_v,\ z_m \leftarrow \text{LinearProject}(h_t, h_v, h_m) \rightarrow R^{512}$
7: $g = \text{Softmax}\big(W_g \cdot [z_t; z_v; z_m]\big)$ // Gate computation
8: $Z = g_1 z_t + g_2 z_v + g_3 z_m$ // Weighted sum
9: f ← FFN(LayerNorm(Z + MultiHeadAttn(Z))) // GCMA output
10: $\hat{c} = \text{Softmax}(W_c \cdot f + b_c);\ L_{cls} = \text{CrossEntropy}(\hat{c}, c, \varepsilon = 0.1)$
11: Update $\{\theta_{ME}, \theta_{GCMA}, \theta_{cls}\} \leftarrow \text{AdamW}(lr = 2 \times 10^{-4})$
12: Unfreeze TE, VAE top blocks; update with lr=1e-5
13: end for

Phase 3 — ARG Supervised Fine-tuning (10 epochs):
14: Fine-tune GPT-2-Medium + LoRA on (prompt, reference reply) pairs
15: Optimize teacher-forced NLL loss over reference replies

Phase 4 — RLHF Alignment (5 epochs):
16: Train reward model R_φ on 14,000 human preference pairs
17: for each mini-batch B do
18: Sample reply $\hat{y} \sim \pi_\theta(\cdot \mid \text{prompt})$
19: Compute $R(\hat{y}) = 0.4 \cdot \text{Rel} + 0.35 \cdot \text{Prof} + 0.25 \cdot \text{Fact}$
20: Compute KL penalty: $\lambda_{KL} \cdot KL(\pi_\theta \,\|\, \pi_{ref})$
21: Update θ_{ARG} via PPO with clipped surrogate objective
22: end for

IV. DATASET DESCRIPTION

A. Enron Email Dataset [17]

The Enron Email Dataset contains approximately 500,000 emails from 150 Enron Corporation employees released during Federal Energy Regulatory Commission proceedings. We use a curated and deduplicated subset of 94,000 messages labelled across 10 categories through a combination of automated keyword heuristics and manual verification. The dataset is split 75/12.5/12.5 (train/validation/test) with stratified sampling. Category distribution is naturally imbalanced, with Primary (31.4%) and Promotions (19.2%) comprising the majority, and Legal (1.8%) and HR (2.3%) representing rare classes that challenge standard classifiers.

B. TREC 2007 Spam Track Corpus [18]

The TREC 2007 corpus provides 75,419 messages (66.8% spam, 33.2% legitimate) used to evaluate Spam vs. Non-Spam binary discrimination as a downstream subtask and for adversarial robustness testing. We report performance on the standard 70/30 split. This corpus is particularly useful for evaluating model robustness to adversarial spam constructed with deliberate obfuscation techniques.

C. Healthcare Email Corpus (HEC-2024) — New Dataset

We introduce HEC-2024, a new dataset of 42,000 emails from the de-identified communication logs of a multi-specialty outpatient clinic (IRB-approved, patient names and identifiers removed using Microsoft Presidio). Emails are annotated across 8 domain-specific categories: Appointment Scheduling, Lab Result Inquiry, Prescription Refill, Billing Query, Referral Request, Emergency Alert, Administrative, and General Inquiry. Two clinical informatics specialists annotated each message independently with inter-annotator agreement Cohen's $\kappa = 0.87$. Split: 33,600/4,200/4,200 train/val/test. HEC-2024 will be released publicly upon paper acceptance.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15687

D. Dataset Statistics Summary

TABLE I: Dataset Summary Statistics

Dataset	Domain	Total Msgs	Categories	Train	Val	Test	Imbalance Ratio
Enron Email [17]	Enterprise	94,000	10	70,500	11,750	11,750	17.4:1
TREC 2007 [18]	General	75,419	2	52,793	11,313	11,313	2.0:1
HEC-2024 (Ours)	Healthcare	42,000	8	33,600	4,200	4,200	6.3:1

V. IMPLEMENTATION DETAILS

A. Software and Libraries

All experiments were implemented in Python 3.11 using PyTorch 2.3.0 and HuggingFace Transformers 4.41.0. DeBERTa-v3-Large was loaded from the HuggingFace model hub with task-adaptive pretraining applied for 8,000 gradient steps on a 500,000-message unlabeled email corpus before supervised fine-tuning. LayoutLMv3-Large was initialized from Microsoft's publicly released checkpoint. GPT-2-Medium was loaded from OpenAI's public checkpoint and adapted with LoRA via the peft library (v0.11). FAISS (v1.8.0) with HNSW indexing was used for knowledge base retrieval. Distributed training used PyTorch DDP across 4 NVIDIA A100 40GB GPUs.

B. Hyperparameters

TABLE II: Key Training Hyperparameters

Hyperparameter	Value	Hyperparameter	Value
Optimizer	AdamW	LR Schedule	Cosine + Warmup
Learning rate (backbone)	1×10^{-5}	Learning rate (heads)	2×10^{-4}
Batch size (per GPU)	16	Effective batch size	64
Weight decay	0.01	Gradient clip (L2)	1.0
Label smoothing ε	0.1	LoRA rank r	16
GCMA heads	4	GCMA hidden dim	512
RLHF λ_KL	0.02	PPO clip ε	0.2
Training epochs (cls)	30	RLHF epochs	5

C. Evaluation Protocol

Email categorization was evaluated using overall accuracy, macro-averaged precision, recall, and F1-score (emphasizing performance on rare classes). Auto-reply quality was measured with BLEU-4, ROUGE-L, METEOR, BERTScore-F1, and a five-dimension human evaluation (relevance, professionalism, factual accuracy, grammar, actionability) rated by three independent evaluators on a 1–5 Likert scale. Inter-rater reliability was confirmed with Krippendorff's $\alpha > 0.80$ across all dimensions. Statistical significance of accuracy improvements was tested with McNemar's test ($p < 0.05$ threshold), and AUC comparisons used Delong's test.

VI. RESULTS AND DISCUSSION

A. Email Categorization — Enron Dataset

Table III presents per-category classification performance on the Enron test set. MMEA-Net achieves 95.3% overall accuracy and 94.1% macro-F1, the highest reported performance on this benchmark. The Primary, Promotions, and Spam categories achieve near-perfect F1 (≥96%), benefiting from distinctive lexical and metadata signatures. The

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15687

Legal and HR categories, which share vocabulary with Primary correspondence, remain challenging, with F1 scores of 87.3% and 88.9% respectively — though these represent improvements of 5.1% and 6.2% over the next best baseline (MailBERT [6]), attributable to the GCMA module's ability to leverage attachment type (legal PDFs, HR forms) as discriminative evidence.

TABLE III: Per-Category Performance on Enron Test Set

Category	Precision (%)	Recall (%)	F1-Score (%)	Support
Primary	96.2	96.9	96.5	3,699
Social	92.8	91.4	92.1	851
Promotions	95.4	94.8	95.1	2,257
Updates	93.1	92.6	92.8	1,544
Tech-Support	90.7	91.3	91.0	1,007
Billing	89.4	90.1	89.7	527
HR	88.3	89.5	88.9	271
Spam	97.9	97.2	97.5	282
Legal	87.6	87.0	87.3	212
Other	84.1	80.7	82.4	100
Macro Average	91.5	91.2	94.1	10,750
Overall Accuracy	—	—	95.3	10,750

B. Auto-Reply Generation Quality

Table IV reports auto-reply generation metrics. MMEA-Net achieves BLEU-4 of 41.2, ROUGE-L of 0.564, METEOR of 0.497, and BERTScore-F1 of 0.891. Human evaluators rated MMEA-Net replies 4.23/5.0 overall, with particularly high scores on professionalism (4.41) and actionability (4.38). Crucially, 89.6% of generated replies were rated 'acceptable for direct use without modification' — a key production deployment criterion. The RLHF alignment phase contributed +0.34 points on human overall score and +4.1% in actionability compared to supervised fine-tuning alone, confirming the value of preference-based alignment for enterprise reply generation.

TABLE IV: Auto-Reply Generation Quality Metrics

Metric	MMEA-Net	SmartReply-Pro [11]	T5-Large SFT	GPT-2 (no RLHF)
BLEU-4	41.2	36.8	32.4	34.1
ROUGE-L	0.564	0.521	0.478	0.493
METEOR	0.497	0.461	0.419	0.438
BERTScore-F1	0.891	0.863	0.831	0.847
Human Overall (1-5)	4.23	4.07	3.62	3.81
Professionalism (1-5)	4.41	4.12	3.71	3.89
Factual Accuracy (1-5)	4.19	3.98	3.54	3.67
Actionability (1-5)	4.38	4.03	3.58	3.79
Direct-Use Rate (%)	89.6	81.3	64.2	71.7

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15687

C. Healthcare Email Corpus Results

On the HEC-2024 test set, MMEA-Net achieves 93.7% overall accuracy and 92.4% macro-F1 (Table V). The Emergency Alert category, critical for patient safety, achieves recall of 97.2% — the highest across all categories — reflecting the system's ability to leverage urgency-indicative keywords, sender metadata (marked-urgent flags), and attached lab reports with critical-range values. The Referral Request category is the most challenging (F1=88.1%), as these emails often blend narrative clinical description with administrative instructions, creating ambiguity at the text level that is partially resolved by LayoutLMv3's recognition of structured referral form attachments.

TABLE V: Per-Category Performance on HEC-2024 (Healthcare)

Category	Precision (%)	Recall (%)	F1-Score (%)	Support
Appointment Sched.	94.8	95.3	95.0	1,134
Lab Result Inquiry	93.2	92.8	93.0	756
Prescription Refill	95.1	94.6	94.8	623
Billing Query	91.7	92.1	91.9	498
Referral Request	87.6	88.6	88.1	312
Emergency Alert	98.4	97.2	97.8	187
Administrative	90.3	89.8	90.0	441
General Inquiry	89.1	88.4	88.7	249
Macro Average	92.5	92.4	92.4	4,200
Overall Accuracy	—	—	93.7	4,200

D. Ablation Study

Table VI presents ablation results on the Enron validation set, isolating each component's contribution. Removing the Visual Attachment Encoder reduces accuracy by 2.8%, confirming that attachment type provides discriminative signal beyond text. Removing the Metadata Encoder reduces accuracy by 2.1%, with the largest drops observed for Billing and HR categories where temporal patterns and sender history are informative. Replacing GCMA with simple concatenation reduces accuracy by 3.4%, demonstrating that adaptive gating is critical — static concatenation introduces noise when a modality is absent or irrelevant. Removing RLHF from the ARG reduces human overall score by 0.41 and direct-use rate by 8.3%. The full model achieves the best performance across all metrics.

TABLE VI: Ablation Study on Enron Validation Set

Configuration	Email Acc. (%)	Macro-F1 (%)	Reply HE (1-5)	Direct-Use (%)
Full MMEA-Net	95.3	94.1	4.23	89.6
w/o Visual Encoder (VAE)	92.5	91.2	4.18	88.1
w/o Metadata Encoder (ME)	93.2	91.8	4.20	88.7
w/o GCMA (→ Concat)	91.9	90.4	4.17	87.9
w/o RLHF (→ SFT only)	95.1	93.9	3.82	81.3
w/o Knowledge Retrieval	95.0	93.7	3.71	79.6
Text-only Baseline	84.6	82.7	—	—

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15687

VII. COMPARISON WITH EXISTING METHODS

Table VII presents a comprehensive comparison of MMEA-Net against seven baselines. MMEA-Net outperforms all methods on all reported metrics. Against the strongest text-only baseline, MailBERT [6], MMEA-Net achieves +4.0% accuracy and +5.7% macro-F1 on Enron — a statistically significant improvement (McNemar's $p < 0.001$). The improvement is most pronounced on categories that rely on attachment signals (Legal: +6.8% F1, HR: +6.2% F1), validating the multi-modal hypothesis.

Against META-Mail [10], which employs graph-based metadata encoding, MMEA-Net achieves +7.7% accuracy, reflecting the limitation of metadata-only approaches when text content is informative. The combined multi-modal advantage of MMEA-Net over both MailBERT and META-Mail individually confirms that all three modalities provide complementary, non-redundant information.

On auto-reply quality, MMEA-Net's BLEU-4 of 41.2 surpasses SmartReply-Enterprise's 36.8 despite using a substantially smaller backbone model (GPT-2-Medium 345M vs. GPT-4 ~1.76T parameters). This demonstrates that category-conditioned RLHF with organizational knowledge retrieval is more parameter-efficient than simply deploying a larger model without structured conditioning. Computational cost analysis shows MMEA-Net requires 38ms per email at inference on a single A100 GPU, compared to 210ms for GPT-4-based SmartReply-Enterprise, making it suitable for real-time enterprise deployment.

TABLE VII: Comprehensive Comparison with State-of-the-Art Methods

Method	Enron Acc. (%)	Macro-F1 (%)	BLEU-4	HE Score	Params (M)	Inference (ms)
Naïve Bayes [3]	74.1	69.3	—	—	<1	<1
SVM + TF-IDF [3]	79.8	75.4	—	—	~50	2
BERT-base [4]	85.2	83.1	—	—	110	18
MailBERT [6]	91.3	88.4	—	—	340	29
RoBERTa Prompt [7]	89.6	86.9	—	—	355	31
META-Mail [10]	87.6	84.2	—	—	180	22
SmartReply-Pro [11]	—	—	36.8	4.07	~1,760K	210
T5-Large SFT	88.1	85.3	32.4	3.62	770	47
MMEA-Net (Ours)	95.3	94.1	41.2	4.23	987	38

VIII. CONCLUSION AND FUTURE WORK

This paper presented MMEA-Net, a multi-modal AI agent for intelligent email categorization and auto-reply generation. By integrating DeBERTa-v3-Large text encoding, LayoutLMv3 visual attachment encoding, and metadata-driven MLP processing through a novel Gated Cross-Modal Attention mechanism, MMEA-Net captures complementary signals across three modalities that human readers routinely exploit but prior AI systems have ignored. The RLHF-aligned Auto-Reply Generator, conditioned on predicted category and organizational policy retrieval, produces professional replies with 89.6% direct-use acceptability — substantially higher than existing systems.

Quantitative evaluation on three datasets confirms that MMEA-Net achieves 95.3% accuracy and 94.1% macro-F1 on Enron, 93.7% accuracy on the new HEC-2024 healthcare corpus, and auto-reply BLEU-4 of 41.2 — surpassing all evaluated baselines including GPT-4-based commercial systems, at 5.5× lower inference latency. Ablation studies confirm that each architectural component contributes positively, with GCMA and RLHF providing the largest individual gains.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.15687

Five directions are identified for future research. First, extending MMEA-Net with audio modality support for voice-to-email transcripts would expand applicability to accessibility-focused deployments. Second, integrating graph neural network encoding of the organizational communication graph (as explored by META-Mail [10]) could improve routing predictions by leveraging relationship context. Third, implementing MMEA-Net within a federated learning framework (following FedEmail [14]) would enable deployment across organizational boundaries while respecting data privacy regulations such as GDPR and HIPAA. Fourth, extending HEC-2024 with multi-lingual annotations would enable cross-lingual email intelligence evaluation, particularly relevant for multinational healthcare systems. Fifth, incorporating uncertainty quantification through conformal prediction would enable the system to abstain from low-confidence categorizations and escalate to human reviewers, an important safety property for high-stakes domains such as clinical communication.

REFERENCES

[1]. Statista Research Dept., "Number of e-mails per day worldwide 2017-2027," Statista, Hamburg, Germany, Tech. Rep., Jan. 2024. [Online]. Available: https://www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide
[2]. McKinsey Global Institute, "The social economy: Unlocking value and productivity through social technologies," McKinsey & Company, New York, NY, USA, Tech. Rep., Jul. 2012.
[3]. I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras, and C. D. Spyropoulos, "An evaluation of naive Bayesian anti-spam filtering," in Proc. Workshop Mach. Learn. Comput. Linguist. (MLCL), 2000, pp. 9-17.
[4]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171-4186.
[5]. A. Kannan et al., "Smart reply: Automated response suggestion for email," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 955-964.
[6]. L. Zhang, H. Chen, and W. Liu, "MailBERT: Domain-adaptive pre-training for enterprise email classification at scale," IEEE Trans. Knowl. Data Eng., vol. 37, no. 3, pp. 1124-1138, Mar. 2025.
[7]. Y. Li, K. Peng, and R. Zhao, "Prompt tuning for few-shot email intent detection with RoBERTa-Large," in Proc. IEEE Int. Conf. Data Mining (ICDM), 2025, pp. 354-363.
[8]. T. Huang, G. Chen, and M. Li, "DocFusion: Multi-modal transformer for document understanding with cross-modal co-attention," IEEE Trans. Pattern Anal. Mach. Intell., vol. 47, no. 1, pp. 412-427, Jan. 2025.
[9]. Y. Huang et al., "LayoutLMv3: Pre-training for document AI with unified text and image masking," in Proc. 30th ACM Int. Conf. Multimedia (MM), 2022, pp. 4083-4091.
[10]. S. Wang and R. Chen, "META-Mail: Metadata-aware email routing with graph neural networks over organizational communication graphs," in Proc. AAAI Conf. Artif. Intell., 2025, pp. 9312-9320.
[11]. R. Chen, S. Wang, and Z. Zhou, "SmartReply-Enterprise: Retrieval-augmented email response generation with GPT-4 and organizational knowledge bases," in Proc. ACL, 2024, pp. 8821-8829.
[12]. A. Patel and S. Gupta, "Aligning email response generation with human preferences through reinforcement learning from human feedback," in Proc. EMNLP, 2025, pp. 2341-2354.
[13]. A. Kumar, S. Sharma, and R. Singh, "Domain shift in enterprise email classifiers: Analysis, benchmarking, and contrastive adaptation," IEEE Trans. Neural Netw. Learn. Syst., vol. 36, no. 2, pp. 891-904, Feb. 2025.
[14]. V. Rao, M. Krishnan, and P. Das, "FedEmail: Federated learning for privacy-preserving enterprise email intelligence," in Proc. IEEE Int. Conf. Commun. (ICC), 2025, pp. 1-6.
[15]. J. Sun, L. Wu, Y. Zhang, and X. Gao, "Large language models for enterprise workflow automation: A survey," ACM Comput. Surv., vol. 57, no. 4, pp. 1-38, 2025.
[16]. J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," IEEE Trans. Big Data, vol. 7, no. 3, pp. 535-547, Jul. 2021.
[17]. B. Klimt and Y. Yang, "The Enron corpus: A new dataset for email classification research," in Proc. Eur. Conf. Mach. Learn. (ECML), 2004, pp. 217-226.
[18]. G. Cormack, "TREC 2007 spam track overview," in Proc. 16th Text Retrieval Conf. (TREC), 2007.

