ANALYSIS OF GENERATIVE MODELS IN AUTOMATED FORMATION OF CORPORATE LETTER TEMPLATES FOR EDUCATIONAL SYSTEMS

https://doi.org/10.70728/tech.v3.i04.007

Axmadaliyev Mansurbek Erkaboy o'g'li
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Department of Information Educational Technologies, Trainee Teacher
axmadaliyevmansur@gmail.com,
Halimov Og'abek Ibodulla o'g'li
Master's student at Tashkent University of Information Technologies named after Muhammad al-Khwarizmi,
halimovogabek85@gmail.com

ABSTRACT The rapid development of artificial intelligence and natural language processing technologies has opened new opportunities for automating document workflows in educational institutions. This study presents a comprehensive comparative analysis of five prominent generative language models - GPT-4, GPT-3.5-Turbo, T5-Large, BERT (fine-tuned), and BLOOM-7B - evaluated on their capacity to generate high-quality corporate letter templates in educational systems. Experiments were conducted on a corpus of 200 authentic institutional letters from Uzbek higher education institutions spanning five letter types. Model performance is assessed using BLEU, ROUGE-L, and F1 metrics alongside a structured human evaluation framework covering fluency, formality, and structural accuracy. Results demonstrate that instruction-tuned large language models significantly outperform encoder-based and smaller generative models, with GPT-4 achieving a BLEU score of 42.3 and a human approval rate of 87%. The study further investigates the impact of prompt engineering strategies, showing that structured few-shot prompts improve GPT-4 performance to a BLEU score of 44.8. Findings provide actionable guidelines for educational institutions considering the deployment of generative AI for administrative document automation.

Keywords: generative models, corporate letter templates, educational systems, natural language processing, GPT-4, T5, BERT, BLOOM, automated document generation, prompt engineering, BLEU, ROUGE

INTRODUCTION

In contemporary educational institutions, the generation and management of official correspondence represents a significant portion of administrative workload. Rectors, department heads, and administrative staff regularly produce a wide variety of formal documents - admission letters, academic notifications, partnership agreements, internal memoranda, and official requests - each requiring strict adherence to institutional style, formal register, and specific structural conventions [1]. The manual composition of such documents is time-consuming, prone to inconsistency, and diverts skilled personnel from

higher-value academic and research activities, particularly in large universities with high volumes of outgoing correspondence.

The administrative burden of document generation is especially pronounced in developing higher education systems undergoing rapid expansion. Universities in Central Asia, including those in Uzbekistan, have experienced significant growth in student enrollment, international partnerships, and interdepartmental communication over the past decade, substantially increasing the volume of required institutional correspondence [2]. This growth has outpaced the development of corresponding administrative infrastructure, creating inefficiencies that negatively affect institutional responsiveness and staff wellbeing.

The emergence of large language models (LLMs) and generative artificial intelligence has introduced promising avenues for automating such tasks. Models capable of understanding context, generating coherent text, and adapting to domain-specific requirements offer the potential to transform document generation from a manual, repetitive task into an intelligent, automated process [3]. However, despite the growing body of research on NLP applications in administrative automation, there remains a notable gap in the literature regarding the systematic comparison of generative models specifically for corporate correspondence generation in educational contexts.

This paper addresses that gap by providing a detailed comparative analysis of five generative models evaluated on a curated corpus of institutional correspondence. The study employs both automatic metrics and a structured human evaluation protocol to provide a comprehensive assessment of model performance across multiple letter types and prompt strategies, and discusses practical implications for deployment in real-world educational settings.

Research Objectives

The primary objectives of this study are: (1) to review the theoretical foundations of generative language models relevant to document generation; (2) to construct a representative evaluation corpus of educational institutional correspondence; (3) to design and validate a multi-dimensional evaluation framework covering both automatic and human assessment; (4) to present comparative experimental results across five selected models and four prompt engineering strategies; and (5) to propose practical guidelines for integrating generative models into educational document management systems.

Scope and Limitations

This study focuses on English-language letter generation to maximize model comparability and align with international publication standards. The evaluation corpus is drawn from Uzbek higher education institutions, providing a domain-specific grounding while acknowledging that cross-institutional and cross-national generalizability may vary. The models evaluated represent a cross-section of current generative architectures; however, the rapidly evolving landscape of LLMs means that newer models released after the study period may exhibit different performance profiles.

Fine-tuning was conducted on T5-Large and BERT only; GPT-4, GPT-3.5-Turbo, and BLOOM-7B were evaluated using API access and prompt-based approaches without additional fine-tuning.

LITERATURE REVIEW AND METHODOLOGY

Literature Review

Research on automated text generation has expanded considerably since the introduction of transformer-based architectures [4]. Vaswani et al. (2017) established the foundational self-attention mechanism that underlies most modern LLMs, enabling models to capture long-range dependencies in text more effectively than preceding recurrent architectures. Devlin et al. (2019) demonstrated the effectiveness of bidirectional pre-training with BERT for a wide range of downstream NLP tasks, establishing masked language modeling as a highly transferable pre-training objective [5].

The subsequent development of decoder-only models - most notably the GPT series by OpenAI - shifted the paradigm toward autoregressive generative applications, enabling models to produce fluent, contextually coherent long-form text [6]. Brown et al. (2020) demonstrated GPT-3's few-shot capabilities for diverse text generation tasks, including formal correspondence, establishing in-context learning as a practical approach for adapting large models to specialized domains without explicit fine-tuning [7]. Raffel et al. (2020) introduced the T5 framework, reformulating all NLP tasks as text-to-text problems and demonstrating strong cross-task transfer learning capabilities [8].

More recently, instruction-tuned models such as InstructGPT and GPT-4 have shown marked improvements in following structured prompts, a property particularly valuable for template-based document generation [9]. Ouyang et al. (2022) demonstrated that reinforcement learning from human feedback (RLHF) substantially improves model alignment with user intent, reducing hallucination and improving output relevance in generation tasks. Scao et al. (2022) presented BLOOM, a multilingual open-access large language model trained on a diverse corpus spanning 46 languages, providing a community alternative to proprietary models for research and deployment in resource-constrained settings [10].

In the domain of administrative document automation, several studies have explored NLP-based approaches for form filling, template completion, and email generation. Maynez et al. (2020) investigated faithfulness and factuality in abstractive summarization, highlighting the challenge of grounding generated text in source documents - a concern directly relevant to institutional correspondence where factual accuracy is critical [11]. Zhang et al. (2020) introduced BERTScore, an evaluation metric leveraging contextual embeddings that correlates more strongly with human judgment than n-gram-based metrics for generation quality assessment [12].

However, existing research has largely focused on general-domain text generation, customer service automation, or healthcare documentation. The specific requirements of educational corporate correspondence - including formal institutional language,

hierarchical address structures, compliance with national and international academic standards, and the need for culturally appropriate salutation and closing conventions - have received comparatively little systematic attention. This study contributes to filling that gap by providing the first domain-specific comparative evaluation focused on educational institutional letter generation.

Corpus Construction

A corpus of 200 authentic corporate letters from Uzbek higher education institutions was collected under a data sharing agreement with five universities, with all personally identifiable information anonymized prior to analysis. The corpus was stratified across five letter categories: official notifications (n=45), partnership requests (n=40), student recommendation letters (n=42), internal memoranda (n=38), and academic inquiry letters (n=35). Letters were collected from the period 2019–2023, ensuring coverage of pre- and post-pandemic administrative communication patterns. Each letter was verified by two independent domain experts for authenticity and representativeness before inclusion.

The corpus was divided into a development set (150 letters) used for prompt design and fine-tuning, and a held-out evaluation set (50 letters, 10 per category) used exclusively for final performance assessment. This separation ensures that evaluation results reflect genuine generalization rather than prompt overfitting. Corpus statistics including average letter length (347 words), vocabulary size (8,240 unique tokens), and formality index scores were documented to characterize the domain.

Model Configurations

Five models were selected to represent distinct points in the current landscape of generative architectures. GPT-4 (OpenAI, March 2023) represents the current state of the art in instruction-tuned large-scale models and was accessed via the OpenAI API with a system prompt establishing the institutional context. GPT-3.5-Turbo (OpenAI, 2022) was included as a cost-effective alternative within the same model family. T5-Large (770M parameters) was fine-tuned for 5 epochs on the 150-letter development corpus using a learning rate of 3×10^{-4} and a batch size of 8 on a single NVIDIA A100 GPU. BERT-base-uncased was adapted for generation by appending a linear language modeling head and fine-tuned on the same corpus. BLOOM-7B (BigScience, 2022) was evaluated in a zero-shot configuration using the Hugging Face Inference API.

Prompt Engineering Framework

A systematic prompt engineering framework was developed and evaluated across four strategies of increasing structure. The unstructured strategy provided only a brief task description without specifying letter components. The semi-structured strategy added explicit fields for letter type and recipient designation. The fully structured strategy incorporated letter type, recipient, institutional context, required sections, and formality level as distinct prompt elements, along with a role assignment instructing the model to act as an experienced university administrative officer. The structured few-shot strategy extended the fully structured prompt with two in-context examples drawn from the development corpus.

All prompts were designed to elicit outputs conforming to a standardized letter structure comprising: institutional header, date, recipient address, formal salutation, subject line, body paragraphs (introduction, main content, call to action), formal closing, and signatory block. This structure was derived from an analysis of the most common structural patterns in the development corpus and validated against institutional style guides from three participating universities.

Evaluation Metrics

Automatic evaluation employed three complementary metrics. BLEU (Papineni et al., 2002) measures n-gram overlap between generated and reference texts, providing a measure of lexical fidelity [13]. ROUGE-L measures the longest common subsequence between generated and reference texts, capturing structural similarity beyond simple n-gram matching. F1-Score was computed as the harmonic mean of ROUGE precision and recall to provide a balanced single-metric summary. Human evaluation was conducted by a panel of five expert evaluators - two university administrative officers, two English language professionals, and one NLP researcher - who rated each output on fluency, formality, and structural accuracy using a 5-point Likert scale. Inter-rater reliability was assessed using Fleiss' kappa, yielding $\kappa = 0.71$, indicating substantial agreement.

RESULTS

Overall Model Performance

The comparative evaluation of the five generative models yielded clear and consistent distinctions in performance across all evaluation dimensions. Table 1 presents the aggregated quantitative results including BLEU, ROUGE-L, F1-Score, and human evaluation approval rates across all 50 held-out letter generation tasks.

Table 1. Overall comparative performance of generative models on corporate letter template generation

Model	BLEU Score	ROUGE-L	F1-Score	Human Eval (%)
GPT-4	42.3	0.61	0.68	87
GPT-3.5-Turbo	38.7	0.57	0.63	79
T5-Large	31.2	0.49	0.54	64
BERT (fine-tuned)	24.5	0.41	0.45	51
BLOOM-7B	29.8	0.46	0.51	60

GPT-4 achieved the highest performance across all metrics (BLEU 42.3, ROUGE-L 0.61, F1 0.68, human approval 87%), reflecting its superior instruction-following capabilities. GPT-3.5-Turbo followed closely (BLEU 38.7, 79% approval), confirming its viability as a cost-effective alternative. T5-Large and BLOOM-7B achieved moderate results of 31.2 and 29.8 BLEU respectively, while BERT performed weakest (BLEU 24.5), consistent with its encoder-only architecture being misaligned with open-ended generation tasks.

Human Evaluation Breakdown

Table 2 presents the disaggregated human evaluation scores across the three assessment dimensions - fluency, formality, and structural accuracy - rated on a 5-point Likert scale. This breakdown reveals important qualitative differences between models that aggregate scores partially obscure.

Table 2. Human evaluation scores by dimension (5-point Likert scale, mean scores)

Model	Fluency	Formality	Structure	Avg
GPT-4	4.6 / 5	4.5 / 5	4.4 / 5	4.5
GPT-3.5-Turbo	4.2 / 5	4.1 / 5	4.0 / 5	4.1
T5-Large	3.5 / 5	3.3 / 5	3.4 / 5	3.4
BERT (fine-tuned)	2.8 / 5	2.6 / 5	2.5 / 5	2.6
BLOOM-7B	3.2 / 5	3.0 / 5	3.1 / 5	3.1

GPT-4 scored highest across all three dimensions (fluency 4.6, formality 4.5, structure 4.4 out of 5), with outputs rarely distinguishable from human-authored letters. GPT-3.5-Turbo followed closely, with its relative weakness in structural accuracy (4.0/5) reflecting occasional omission of the institutional header block. T5-Large showed a notable gap between fluency (3.5/5) and formality (3.3/5), with evaluators flagging casual phrasing in closing sections. BERT received the lowest scores overall, with structural accuracy (2.5/5) its most significant weakness.

Performance by Letter Type

Table 3 presents GPT-4 performance disaggregated by letter type, providing insight into which correspondence categories benefit most from generative automation. GPT-4 was selected for this analysis as the highest-performing model; equivalent breakdowns for other models showed proportionally similar patterns.

Table 3. GPT-4 performance by letter type

Letter Type	BLEU	ROUGE-L	Human Eval (%)
Official Notification	44.1	0.63	89
Partnership Request	41.8	0.60	86
Recommendation Letter	43.2	0.62	88
Internal Memorandum	40.5	0.59	84
Academic Inquiry	42.7	0.61	87

Official notifications yielded the highest results (BLEU 44.1, 89% approval), reflecting their highly formulaic structure. Recommendation letters also performed strongly (BLEU 43.2, 88% approval). Internal memoranda showed the lowest scores within GPT-4 results (BLEU 40.5, 84% approval), as the model occasionally introduced unnecessary formal preamble not typical of internal communications. These findings confirm that explicitly specifying letter type in deployment prompts is essential for appropriate register calibration.

Error Analysis

Human evaluators identified five recurring failure categories across all models: incorrect salutation structure, inconsistent institutional referencing, missing closing formulae, inappropriate informal register, and structural inconsistency. GPT-4 demonstrated the lowest overall error rates, with missing closing formulae its most frequent failure (7% of outputs) - easily addressed by explicitly including the closing block in the prompt template. Incorrect salutation appeared in only 6% of GPT-4 outputs, limited to cases where institutional hierarchy was unspecified in the prompt.

BERT exhibited the highest error rates, with wrong salutation (31%) and structural inconsistency (25%) as primary failure modes, confirming its architectural misalignment with sequential generation tasks. T5-Large errors concentrated in salutation construction (18%) and informal register (12%), reflecting its broad pre-training distribution skewing toward casual web text. BLOOM-7B showed moderate errors similar to T5-Large, with inconsistent institutional referencing (12%) as its most distinctive weakness.

Impact of Prompt Engineering Strategy

GPT-4 was evaluated across four prompt strategies of increasing structure to isolate the contribution of prompt design to output quality. Unstructured prompts yielded a BLEU score of 31.4 and human approval of 62% - comparable to fine-tuned T5-Large - demonstrating that even state-of-the-art models underperform significantly without adequate prompt scaffolding. The semi-structured strategy (adding letter type and recipient designation) raised BLEU to 37.9 and approval to 76%, while the fully structured strategy (incorporating role assignment, required sections, and formality level) achieved BLEU 42.3 and approval 87%.

The few-shot strategy - extending fully structured prompts with two in-context examples - achieved the highest performance (BLEU 44.8, approval 91%), demonstrating that example-based prompting efficiently communicates institutional style conventions that are difficult to specify exhaustively in a system prompt. The 13.4 BLEU-point gap between unstructured and few-shot prompting strongly supports the development of institution-specific prompt libraries as a priority investment for any educational deployment of generative models.

DISCUSSION

The results of this study confirm that large-scale instruction-tuned generative models, particularly GPT-4, represent the most capable current solutions for automated corporate letter generation in educational systems. Their ability to interpret nuanced prompts, maintain formal register across extended outputs, and adapt to varied letter types reflects the advantages of scale and reinforcement learning from human feedback (RLHF) in shaping model behavior for structured document tasks [9].

The prompt engineering findings carry particularly significant practical implications. The 13.4 BLEU-point gap between unstructured prompting (BLEU 31.4) and few-shot prompting (BLEU 44.8) of the same GPT-4 model demonstrates that prompt quality constitutes a major determinant of output quality - comparable in magnitude to the entire performance gap between GPT-4 and T5-Large under optimal conditions.

Notably, GPT-4 operating under unstructured prompts performed only marginally above fine-tuned T5-Large, suggesting that without structured prompting, the cost and capability advantages of large proprietary models are substantially eroded. This finding implies that educational institutions should treat prompt library development as a first-order priority in any generative AI deployment initiative, investing in systematic prompt design and validation before committing to a specific model architecture or API subscription.

Data privacy remains a critical consideration for deployment. Institutional correspondence may contain sensitive personal or organizational information, and cloud-based API services require careful compliance with applicable data protection frameworks. Educational institutions in Uzbekistan must consider alignment with national data localization requirements under the Law on Personal Data, in addition to any bilateral agreements governing data processing with non-resident service providers. For institutions with stringent data sovereignty requirements, the performance of fine-tuned T5-Large - while lower than GPT-4 - may represent an acceptable trade-off enabling fully on-premise deployment.

The finding that BLOOM-7B achieves competitive zero-shot performance despite its significantly smaller parameter count relative to GPT-4 is noteworthy, and suggests that continued development of open-access multilingual models may yield viable deployment alternatives as model quality improves. Future work should examine BLOOM's performance after domain-specific fine-tuning, which was not conducted in the present study due to computational constraints.

The relatively weak performance of BERT-based generation further highlights the importance of architectural alignment between model design and task requirements. While BERT remains a powerful tool for classification, named entity recognition, and information extraction tasks, its encoder-only architecture creates fundamental limitations for open-ended generation. Educational institutions should avoid repurposing it for correspondence automation without substantial architectural modification, such as the addition of a cross-attention decoder as in encoder-decoder models.

CONCLUSION

This study has presented a systematic comparative analysis of five generative language models - GPT-4, GPT-3.5-Turbo, T5-Large, BERT (fine-tuned), and BLOOM-7B - for the automated generation of corporate letter templates in educational institutions. Experiments on a curated 200-letter corpus drawn from Uzbek higher education institutions demonstrate that instruction-tuned large language models, particularly GPT-4, currently offer the highest performance for this task, achieving a BLEU score of 42.3 and a human evaluation approval rate of 87% under fully structured prompting, rising to 44.8 and 91% respectively with few-shot prompting. Smaller open-source models such as T5-Large and BLOOM-7B provide viable alternatives when domain-specific fine-tuning is applied or computational and data sovereignty constraints preclude the use of proprietary cloud APIs.

The study makes three primary contributions to the literature. First, it provides the first domain-specific comparative evaluation of generative models for educational institutional correspondence generation, establishing baseline performance figures for future research. Second, it demonstrates the substantial impact of prompt engineering strategy on output quality, with structured few-shot prompting improving performance by over 13 BLEU points relative to unstructured prompting. Third, it provides a validated multi-dimensional human evaluation framework - covering fluency, formality, and structural accuracy - that can be adopted and extended by future research in adjacent domains.

Key practical recommendations for educational institutions include: prioritizing structured prompt library development before model selection; evaluating fine-tuned open-source models where data privacy requirements preclude cloud API usage; specifying letter type explicitly in all deployment prompts to enable appropriate register calibration; and establishing data governance protocols and inter-rater evaluation pipelines prior to full deployment. Future research should explore multilingual generative models capable of producing high-quality correspondence in Uzbek and Russian, the integration of retrieval-augmented generation (RAG) to ground outputs in institutional templates and regulatory documents, and longitudinal studies assessing the impact of AI-assisted correspondence on administrative efficiency and staff workload in real deployment settings.

REFERENCES

[1] Ministry of Higher Education, Science and Innovation of the Republic of Uzbekistan. (2022). Regulations on official correspondence in higher education institutions. Tashkent: MHESI Press.

[2] UNESCO Institute for Statistics. (2022). Higher education enrollment trends in Central Asia 2010-2022. UNESCO.

[3] Bommasani, R., Hudson, D. A., Aditi, E., et al. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171-4186.

[6] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8).

[7] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

[8] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.

[9] OpenAI. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774.

[10] Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilic, S., Hesslow, D., et al. (2022). BLOOM: A 176B-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100.

[11] Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. Proceedings of ACL 2020, 1906–1919.

[12] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. Proceedings of ICLR 2020.

[13] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. Proceedings of ACL 2002, 311–318.

[14] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.