Comparative Evaluation of AI Models for Business Email Writing: Assessing Formality, Readability, and Performance

Shivangi Narayan
Department of Computer Science
School of Engineering and Sciences
SRM University AP
Guntur, India
shivangi_narayan@srmap.edu.in

Pranathi Jayanthi
Department of Computer Science
School of Engineering and Sciences
SRM University AP
Guntur, India
pranathi_j@srmap.edu.in

Ayush Kumar
Department of Computer Science
School of Engineering and Sciences
SRM University AP
Guntur, India
ayush_kumar@srmap.edu.in

Nazeer Ahmed
Department of Computer Science
School of Engineering and Sciences
SRM University AP
Guntur, India
nazeer_ahmed@srmap.edu.in

Ajay Bhardwaj
Department of Computer Science
School of Engineering and Sciences
SRM University AP
Guntur, India
ajay.b@srmap.edu.in

***Abstract*—Large Language Models (LLMs) have revolutionized business communication with automated email writing, improved efficiency, and personalization. This paper presents a comparative study of four widely used AI-powered LLMs in business email writing considering the following parameters: formality, readability, grammatical correctness, and word limit. This research compares models such as OpenAI-GPT-4o-mini, Google-Gemini-2.0-Flash, Claude-3.7-Sonnet, and Meta-Llama3-70b in producing professional business emails that fit various scenarios. This work quantitatively compares AI-produced emails by adopting automated scoring. Based on our findings, Google-Gemini-2.0-Flash leads its peers in producing professional, polished, and grammatically correct business communication. This research aims to help professionals choose the most appropriate AI LLM to write effective and contextually relevant emails.**

***Index Terms*—Artificial intelligence, large language models, natural language processing, e-mail writing, formality, readability.**

I. Introduction

Effective formal communication is really important across the corporate, academic, and professional sectors. Emails can be used to create connections and make decisions in the digital world. Clear and formal written emails are necessity for coordinating teams and closing business deals. These communications set the tone for interactions and create the first impression of the professionalism and reputation of an organization. As Generative AI has become freely available in recent years, people have started using AI to write emails [1]. Large language models (LLMs) such as ChatGPT and Gemini etc. can generate text that closely mirrors human thought. These models quickly understand context and nuance and can serve as reliable digital writing assistants. These LLMs can handle various formal scenarios from producing daily administrative messages to composing sophisticated persuasive proposals. So, they can improve both efficiency and consistency in communication.

Amidst this technological progress, the most prominent question that arises is: which of these LLMs can consistently generate formal emails that can meet professional and academic standards? To answer this question, our research embarks on a detailed comparative analysis of GPT, Gemini, Claude, and Llama. Our central interest is to examine how each model responds to the different scenarios while maintaining clarity, tone, and appropriate context. Our assessment goes beyond mere functionality. We would like to understand how these models manage to incorporate the human element that is so critical to formal messaging. Our research examines the following scenarios:

1) *Routine Communications:* How do these models perform when creating correct and clear routine emails?
2) *Sensitive Messaging:* How do these tools handle informing bad news or correcting errors? How do they create these tough tactful messages?
3) *Persuasive Writing:* Can such LLMs generate persuasive emails that can actually affect decision-making? How well do they use indirect approaches essential in high-stakes negotiations? [2]

Applications of our findings will be very useful for organizations, professionals, educators and students. With the use of AI tools in writing, not only is time saved but also anxiety and fear while writing formally are avoided. Since there has been

a rise in the use of digital tools for official communication, it has become imperative that the AI models uphold the corporate values and remain productive. For frequent email writers, it is crucial to know the strengths and limitations of these LLMs. This will enable them to write their best with the assistance of AI. By comparing performance in a systematic way, we will provide insights into the potential and limitation of existing AI-driven writing tools. This paper is about using AI optimally in formal writing. Through a comparative analysis of GPT, Gemini, Claude and Llama, we hope to get insights that will help all utilise the power of AI while retaining human elements that makes formal writing effective.

Organization: Section II provides the literature survey, while Section III provides the methodology. In Section IV, results and analysis are provided, while Section V provides the conclusion.

II. Literature Survey

A. The Role of Effective Communication in Business

Being able to communicate effectively can determine the success of your organization as it affects several aspects of business. In a recent work [3], authors Valiyeva & Thomas emphasize that good communication is key to the satisfaction of employees, making decisions and serving customers. These researches outline the inherent value of communication to business, and the setting for an inquiry into how AI technologies may be utilized to support such critical processes. In the same way, the work in [4] note the importance of communication in trust building and cooperation within organizations. These researches highlight the inherent significance of communication to business, laying the groundwork for the examination of how AI technologies can reinforce these vital processes.

B. The Emergence of AI in Business Communication

The application of Artificial Intelligence (AI) technologies, particularly large language models (LLMs), is transforming business communication. Getchell et al. in [5] propose the AI roles framework for the detection of five such broad roles of AI in business communication: Tool, Assistant, Monitor, Coach, and Teammate. The proposes model gives the whole picture of how AI can be utilized to enable different aspects of business communication. Recently, there have been tremendous developments in LLMs, and these have contributed significantly to business writing. These models, developed with the help of large amounts of data, have amazing potential to generate formal material across all modes, ranging from emails to blogs and social media posts. The development of newer versions like ChatGPT, Gemini, and so on has also broadened the usage of AI for business communication.

C. Applications of AI in Business Writing

1) Email Generation and Automation

AI-based email generators are leading the trend last few years because of their capability to generate automated messages, compose personalized content, and maintain compliance in communications. These are NLP and machine learning algorithm based applications that mimic human language. AI email generator in addition to large business will be helpful to small groups or low-resource businesses since they can automate small tasks while allocating human workforce to complex operations.

2) Customer Service and Virtual Assistants

The work in [6] speaks about how AI applications like ChatGPT can boost productivity to a great extent and improve collaborative group work. This results in better customer service and faster problem-solving. Similarly, in another work [7], authors also speak about ChatGPT's role in customer service and virtual support, with a focus on providing the capability to develop more interactive interactions with the customers.

3) Content Optimization and SEO

LLMs are also increasingly being utilized to optimize content for search engines by suggesting suitable keywords, meta description optimization, and content organization for better discoverability. This functionality enables businesses to boost their web presence and interaction.

4) Translation and Localization

Machine translation software provides context-sensitive and accurate translations that retain the original message's intent and tone. This is critical for global business organizations so they can deliver culture-specific content.

D. AI-Generated Content Evaluation

Recent studies quantify the performance of AI-written content in corporate communication. In [8], authors provide a framework for quantifying AI-written emails, while the work in [9] investigates the stylistic features of ChatGPT-4-generated answers in corporate correspondence. The comparative study done by Mnasri & Jovic in [10] talks about the need for an examination of LLM writing skills for business emails. Their study compares the performance of four LLMs (ChatGPT 3.5, Bard, Bing Chat, and Llama 2.0) for different types of business emails. Few of the prominent findings in these tests are:

1) *Strengths:* LLMs are found to be proficient in the use of language, clarity, and pertinence in business communication.
2) *Limitations:* AI-generated content struggles with affective context, tactfulness to the audience, and strict adherence to precise email formats.
3) *Variability:* Due to its formulaic application in business email, there are still vast differences across the output of different LLMs.

E. Challenges and Ethical Considerations

While there are numerous beneficial attributes of business communication via AI, there have been some issues and ethical concerns experienced:

1) Lack of Emotional Intelligence: The recent work [11] surveys that AI content is less emotionally intelligent

and empathetic as required in advanced communication contexts.
2) Authenticity and Trust: People are skeptical about whether AI generated messages are genuine and how they might influence trust in business relationships.
3) Data Privacy and Security: Use of AI in communication raises grave issues of data security and privacy, particularly while handling sensitive business information.
4) Bias and Fairness: AI systems can reinforce or even amplify existing biases if not properly designed and controlled, and this can cause discriminatory messages.
5) Transparency and Accountability: AI system decision-making is opaque, and therefore issues of accountability and transparency are encountered in business communication.

III. Methodology

A. Research Objectives and Scope

The purpose of this study is to evaluate and compare the performance of four of the leading large language models (LLMs) currently in existence, namely OpenAI-GPT-4o-mini, Google-Gemini-2.0-Flash, Claude-3-7-sonnet-20250219, and Meta-llama3-70b-8192, at generating highly professional, polished, and well-crafted formal business emails for specified business scenarios. The final goal is to determine which model generates the best-performing emails, as evaluated based on given requirements such as upholding corporate protocol, clarity, brevity, and overall professionalism of tone.

B. Generation and Preparing of Data

A prompt-based email generation mechanism is employed. Specifically, a structured prompt was employed to generate business email responses in formal tone by each LLM. The prompt set a corporate environment with strict guidelines for providing similarity and adherence to business etiquette. The prompt required a fixed set of recipient (e.g. "Alex Carter"), a predefined signature block (e.g. "John Doe, Senior Manager, ABC Corp."), and word range constraints (150-200 words). The prompt clearly asked for professional tone, simple lexicon, and proper transitions across mail sections. Following LLM models were employed in creating the email datasets:

- OpenAI - gpt-4o-mini
- Google - gemini-2.0-flash
- Claude - claude-3-7-sonnet-20250219
- Meta - llama3-70b-8192

Specifically, the prompt used is: Generate a highly professional, polished, and well-structured formal email for the following business scenario: scenario. The email should be between 150-200 words and must follow strict corporate etiquette. Ensuring that:

- The subject Line is concise, engaging, and relevant.
- The email begins with a polite and professional greeting (using the fixed recipient name: Alex Carter → e.g., 'Dear Alex Carter,').
- The introduction smoothly sets the context in a clear and engaging manner.
- The main Body conveys the core message professionally while maintaining clarity and conciseness.
- The closing Statement includes a clear next step or a polite call to action.
- The signature is consistent across all emails and should be: John Doe, Senior Manager, ABC Corp.

Furthermore, to improve the dataset quality, the following unbreakable guidelines are followed:

- "Alex Carter" must remain the recipient in every single email—no variations.
- The subject and body are the ONLY variables—no extra names, senders, or changes.
- The tone must be formal, polished, and business-appropriate (no casual phrases).
- Keep language clear, professional, and direct—no unnecessary fluff.
- Express gratitude or appreciation where relevant (e.g., "Thank you for your time and consideration.").
- Maintain seamless transitions between sections for natural readability.
- Emails should sound like they were written by a real business professional—not AI-generated.
- Your goal: Make this email feel like it was crafted by a top corporate executive—precise, polished, and persuasive.

Compilation of the Datasets: With each mail generated by the LLMs, a JSON document was used to save the produced emails. For each JSON submission, the output email text using the model came as a reaction to the formalized prompt. The business context varied based on the email created.
Data Preprocessing: JSON files were read using Python's json library. Each JSON record was read and extracted for the email text analysis. Only the created email, and not business context, was analyzed.

C. Evaluation Metrics and Procedures

The evaluation structure was set up to evaluate the quality of generated emails on four essential aspects:

1) *Formality:* A lexical analysis technique was utilized. Formal and informal word lists were pre-determined. The number of informal words was subtracted from the number of formal words to obtain a formality score.
   Implementation: Spacy library was utilized for lexical analysis and tokenization.
2) *Readability:* Flesch-Kincaid Grade-Level score served as the metric of readability [12].
   Implementation: The Flesch-Kincaid Grade Level was computed utilizing the textstat library.
   Interpretation: Lower grade level means easier to read.
3) *Grammar Quality:* Grammar mistakes were identified by the LanguageTool API [13].
   Implementation: The language_tool_python library was utilized for communicating with the LanguageTool API.
   Metric: The count of identified grammar errors was utilized as the grammar quality score. Lower score implies higher grammar quality.

4) *Conciseness:* Word count was utilized as the measure of shortness.
   Implementation: The email content was dissected into words, and the number of words of the list obtained was calculated.
   Metric: The sum of the number of words in the email.

D. Data Analysis

The very first step performed in data analysis is the aggregation of scores in which the scores for each metric for each model were averaging the scores for all emails in the provided dataset.

- Score normalization: To facilitate a fair comparison between measures of varying scales, scores were normalized using min-max normalization.
  Normalized score = (Value - Minimum Value) / (Maximum Value - Minimum Value)
- Weighted scoring: Weighted scoring was used to facilitate balanced scoring. Weights were assigned to every measure based on how important they were perceived to be. In this work, these values are:
  1) Formality: 0.3
  2) Readability: 0.3
  3) Grammar Quality: -0.3 (negative weight since less error is better)
  4) Conciseness: 0.1

  The normalized scores were also multiplied by their corresponding weights, and their values summed to get a final weighted score for each model.
- Statistical analysis: Mean of each measurement was computed for each model. Minimum value and maximum value of each measurement across all models were used to normalize the data.

E. Parameter Calculation and visualization

To delineate the metric-Specific Comparisons, Bar charts were employed to show the performance of each model on each metric. The parameters are calculated as follows:

$$\text{Formality Score} = \max(0, \text{formal word count} - \text{informal word count}) \quad (1)$$

$$\text{FK Grade Level} = 0.39 \times \left(\frac{\text{Total Words}}{\text{Total Sentences}}\right) + 11.8 \times \left(\frac{\text{Total Syllables}}{\text{Total Words}}\right) - 15.59. \quad (2)$$

$$\text{Final Score} = (0.3 \times \text{Formality}) + (0.3 \times \text{Readability}) - (0.3 \times \text{Grammar Errors}) + (0.1 \times \text{Conciseness}) \quad (3)$$

F. Implementation Tools

- Python 3.x
- Libraries: json, os, textstat, spacy, matplotlib.pyplot, numpy, collections, language_tool_python.

TABLE I
COMPARISON OF LANGUAGE MODEL PERFORMANCE METRICS

Model	Formality Score	Readability Score	Grammar Issues	Word Count
OpenAI-GPT-4o-mini	0.01	11.33	3.20	161.20
Google-Gemini-2.0-Flash	0.02	12.08	0.30	141.37
Claude-3.7-Sonnet	0.01	13.31	2.28	175.02
Meta-Llama3-70b	0.00	11.66	0.16	153.72

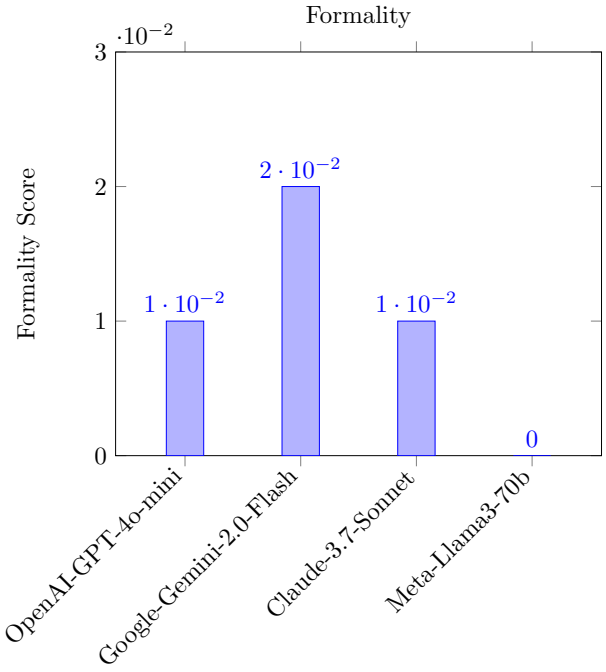

Fig. 1. Formality scores of different AI models

IV. RESULTS AND ANALYSIS

A. Performance Measures

AI models' performance was measured based on professionalism in the email language, clarity, grammatical correctness, and word count compliance. The obtained results are shown in Table 1.

B. Email Formality Analysis

The highest formality score, a value of 0.02, belonged to Google-Gemini-2.0-Flash, which indicated a uniform professional tone. The next were OpenAI-GPT-4o-mini and Claude-3.7-Sonnet with a formality score of 0.01. This shows that they are formal but less than Gemini. The lowest formality score of 0.00 belongs to Meta-Llama3-70b, which means that the emails it generated contained more informal language.

C. Readability Analysis

Figure 2 depicts the readability scores of all LLM Models. It shows that Claude-3.7-Sonnet reported the highest readability

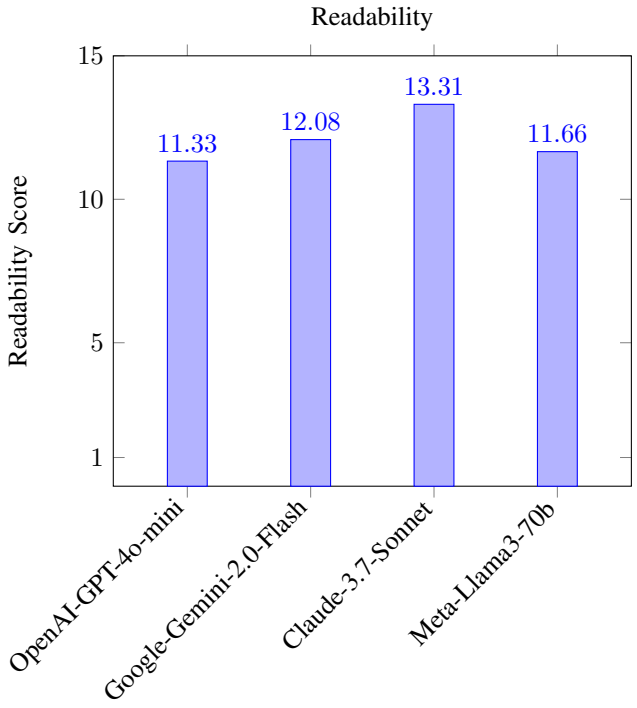

Fig. 2. Readability scores of different AI models

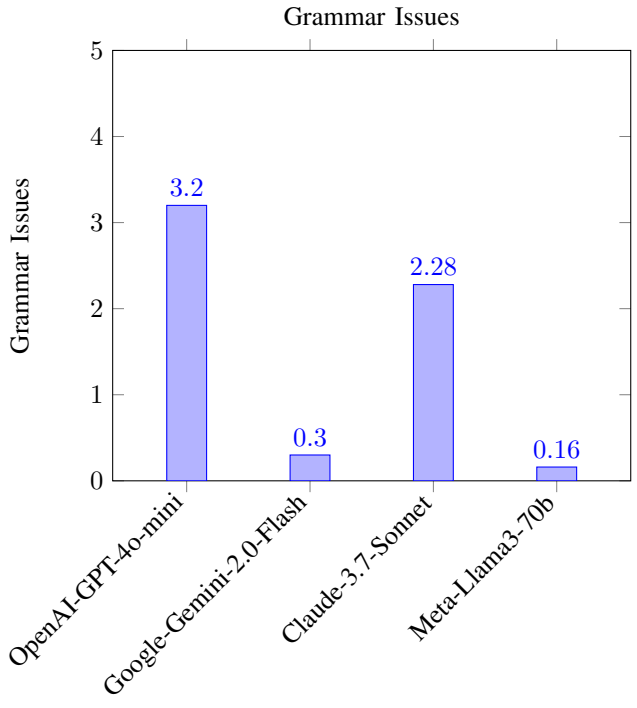

Fig. 3. Grammar issues in different AI models

score with a value of 13.31. This indicates higher clarity and well-written sentences. While Google-Gemini-2.0-Flash also ranked high with a value of 12.08, and thus it is appropriate for business communication. OpenAI-GPT-4o-mini and Meta-Llama3-70b has a score of 11.33 and 11.66, respectively, generated less readable emails than the best-performing models. This indicates that their sentences were too complex and difficult to understand.

D. Grammar Accuracy

Figure 3 depicts the grammar accuracy of various LLMs. Meta-Llama3-70b has a grammar score of 0.16 which signifies it recorded the least number of errors with highest accuracy in sentence formation. Google-Gemini-2.0-Flash also scored good in grammatical accuracy with a meagre 0.30 grammar errors per email, affirming its professional orientation. The Claude-3.7-Sonnet and OpenAI-GPT-4o-mini has a score of 2.28 and 3.20, respectively. It means they are more grammatically error-prone, and thus less credible as far as formal business letters are concerned.

E. Word Count Compliance

Figure 4 depicts the word count compliance of various reported LLMs. Google-Gemini-2.0-Flash and Meta-Llama3-70b hava a value of 141.37 words and 153.72 words, respectively, shows they adhered well to the word limit. OpenAI-GPT-4o-mini (161.20 words) and Claude-3.7-Sonnet (175.02 words) went above the word mark.

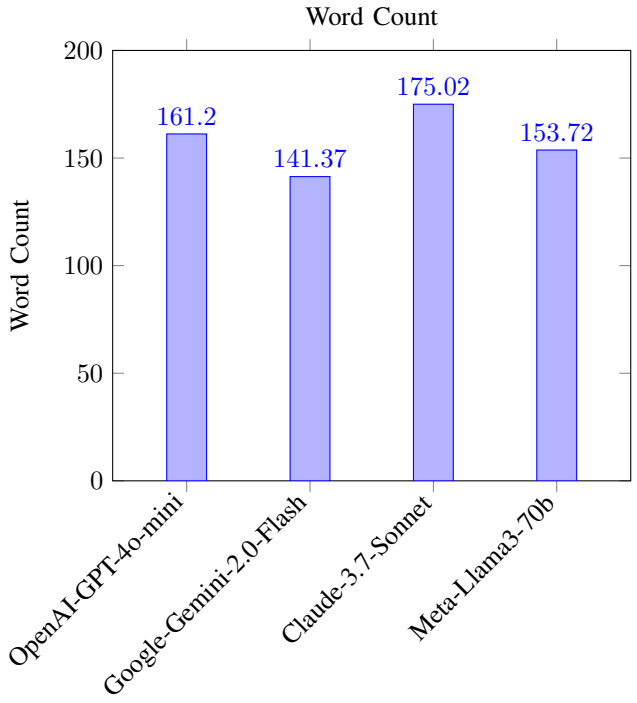

Fig. 4. Word count of AI-generated emails

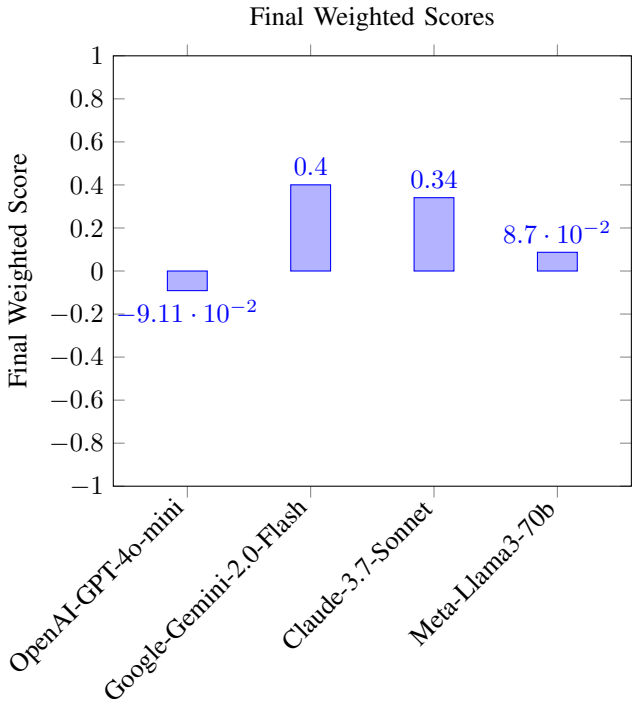

Fig. 5. Final Weighted Scores of AI Models

F. Final Weighted Scores

The above discussed weighted scoring system was used to rank the models according to their performance in all the metrics:

Model	Final Weighted Score
Google-Gemini-2.0-Flash	**0.4005**
Claude-3.7-Sonnet	0.3408
Meta-Llama3-70b	0.0870
OpenAI-GPT-4o-mini	-0.0911

G. Best Performing Model

According to the final weighted scores Google-Gemini-2.0-Flash turns to be the best performing AI model for writing formal business emails. It had the highest formality score, the least number of grammatical errors, and a good readability score, making it the best model to use in professional communication. Although Claude-3.7-Sonnet was well readable, it contained more grammatical errors and was wordy. Meta-Llama3-70b generated grammatically correct text but with less formal tone. OpenAI's GPT 4o-mini performed the worst. It had the most grammatical errors and the lowest readability scores. Its tone is unnecessarily formal and complex. It also uses very long sentences. These results are very meaningful in selecting the right AI model for writing business emails, where professionalism, conciseness, and grammatical accuracy in communication must be guaranteed.

V. Conclusion

In this paper, we compare four leading large language models to write business emails. The comparative study highlights the strengths and weaknesses of these leading LLMs while crafting the business emails. By evaluating formality, readability, grammatical correctness, and word limit, we provide a data-driven analysis of their effectiveness. Our findings indicate that Google-Gemini-2.0-Flash outperforms its counterparts in generating professional and polished emails. Among the four models that we evaluated, the final performance was: Google-Gemini-2.0-Flash, Claude-3.7-Sonnet, Meta-Llama3-70b, and OpenAI-GPT-4o-mini. This research serves as a valuable guide for professionals seeking the most suitable LLM powered by artificial intelligence (AI) to craft effective and contextually relevant business communication. Future work can explore additional evaluation metrics and industry-specific customization to further enhance AI-assisted email writing.

References

[1] P. Gryka, K. Gradoń, M. Kozłowski, M. Kutyła, and A. Janicki, "Detection of ai-generated emails-a case study," in *Proc. of the 19th International Conference on Availability, Reliability and Security*, 2024, pp. 1–8.

[2] A. Rogiers, S. Noels, M. Buyl, and T. De Bie, "Persuasion with large language models: a survey," *arXiv preprint arXiv:2411.06837*, 2024.

[3] A. Valiyeva and B. J. Thomas, "Successful organizational business communication and its impact on business performance: An intra-and inter-organizational perspective," *Journal of Accounting, Business and Finance Research*, vol. 15, no. 2, pp. 83–91, 2022.

[4] C. T. Romm and N. Pliskin, "Toward a virtual politicking model," *Communications of the ACM*, vol. 40, no. 11, pp. 95–100, 1997.

[5] K. M. Getchell, S. Carradini, P. W. Cardon, C. Fleischmann, H. Ma, J. Aritz, and J. Stapp, "Artificial intelligence in business communication: The changing landscape of research and teaching," *Business and Professional Communication Quarterly*, vol. 85, no. 1, pp. 7–33, 2022.

[6] H. Alshurafat, "The usefulness and challenges of chatbots for accounting professionals: Application on ChatGPT," *Available at SSRN 4345921*, 2023.

[7] A. S. George and A. H. George, "A review of ChatGPT AI's impact on several business sectors," *Partners universal international innovation journal*, vol. 1, no. 1, pp. 9–23, 2023.

[8] Y. Sahari, A. M. T. Al-Kadi, and J. K. M. Ali, "A cross sectional study of ChatGPT in translation: Magnitude of use, attitudes, and uncertainties," *Journal of Psycholinguistic Research*, vol. 52, no. 6, pp. 2937–2954, 2023.

[9] M. A. AlAfnan and S. F. MohdZuki, "Do artificial intelligence chatbots have a writing style? an investigation into the stylistic features of ChatGPT-4," *Journal of Artificial intelligence and technology*, vol. 3, no. 3, pp. 85–94, 2023.

[10] S. Mnasri and M. Jovic, "On the need to explicitize the unstated argument in cancer research: an ethnography on scientific argumentation," *Humanities and Social Sciences Communications*, vol. 10, no. 1, pp. 1–9, 2023.

[11] P. Cardon, C. Fleischmann, J. Aritz, M. Logemann, and J. Heidewald, "The challenges and opportunities of AI-assisted writing: Developing AI literacy for the AI age," *Business and Professional Communication Quarterly*, vol. 86, no. 3, pp. 257–295, 2023.

[12] M. Solnyshkina, R. Zamaletdinov, L. Gorodetskaya, and A. Gabitov, "Evaluating text complexity and flesch-kincaid grade level," *Journal of social studies education research*, vol. 8, no. 3, pp. 238–248, 2017.

[13] K. Eker, M. K. Pehlivanoğlu, A. G. Eker, M. A. Syakura, and N. Duru, "A comparison of grammatical error correction models in english writing," in *Proc. of 8th International Conference on Computer Science and Engineering (UBMK)*. IEEE, 2023, pp. 218–223.