

Email subjects generation with large language models: GPT-3.5, PaLM 2, and BERT

Soumaya Loukili, Abdelhadi Fennan, Lotfi Elaachak

Data and Intelligent Systems Team FSTT, Abdelmalek Essaadi University, Tetouan, Morocco

Article Info

Article history:

Received Feb 29, 2024

Revised Mar 20, 2024

Accepted Apr 22, 2024

Keywords:

Artificial intelligence

Deep learning

Digital marketing

GPT

Large language models

ABSTRACT

In order to enhance marketing efforts and improve the performance of marketing campaigns, the effectiveness of language generation models needs to be evaluated. This study examines the performance of large language models (LLMs), namely GPT-3.5, PaLM 2, and bidirectional encoder representations from transformers (BERT), in generating email subjects for advertising campaigns. By comparing their results, the authors evaluate the efficacy of these models in enhancing marketing efforts. The objective is to explore how LLMs contribute to creating compelling email subject lines and improving opening rates and campaign performance, which gives us an insight into the impact of these models in digital marketing. In this paper, the authors first go over the different types of language models and the differences between them, before giving an overview of the most popular ones that will be used in the study, such as GPT-3.5, PaLM 2, and BERT. This study assesses the relevance, engagement, and uniqueness of GPT-3.5, PaLM 2, and BERT by training and fine-tuning them on marketing texts. The findings provide insights into the major positive impact of artificial intelligence (AI) on digital marketing, enabling informed decision-making for AI-driven email marketing strategies.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Soumaya Loukili

Data and Intelligent Systems Team FSTT, Abdelmalek Essaadi University

Tetouan, Morocco

Email: soumaya.loukili1@etu.uae.ac.ma

1. INTRODUCTION

Over the last few years, digital marketing has been significantly impacted by artificial intelligence (AI) [1], completely altering interactions between companies and their customers. Integration of large language models (LLMs) into content creation techniques has been one of the most recent and significant developments in this field [2]. With its astounding 175 billion characteristics [3], LLMs like generative pre-trained transformer 3.5 (GPT-3.5) have ushered in a new era of content creation. These models have demonstrated their ability to produce content of unprecedented quality and contextual relevance on a massive scale. For instance, GPT-3.5 can generate text that resembles human writing across a wide range of topics and genres, including posts for social media, product descriptions, and more. With the use of this capacity, marketers can efficiently create massive amounts of content while saving time and resources and maintaining a consistent brand voice.

LLMs can become essential to the optimization of email marketing campaigns, which can lead to a significant increase in open and click-through rates. This can be done by utilizing the strength of these models to create captivating email subject lines and interesting content. If A/B testing studies demonstrate that emails with subject lines created by LLMs can perform better than emails with subject lines written by humans, generating a rise in open rates, it will highly improve email marketing's efficiency while also freeing up marketing teams to concentrate on other tactical facets of their campaigns. The objective of this study is twofold. Assessing the

performance of different language model-based approaches in marketing, specifically email marketing, is the first part. It will allow for a comparison between the LLM-generated subjects and those created by human efforts. Second, the intention is to execute the said comparison based on key metrics, such as opening rate and other language generation metrics. By conducting this analysis, the authors aim to gain insights into the effective use of LLM in improving email marketing campaigns performance.

This paper is structured as follows: section 2 discusses the key concepts of language modeling, elucidating its definition, various types, and tracing its evolution over time, along with introducing the key focus of this study: three prominent large language models, PaLM 2, GPT-3.5, and bidirectional encoder representations from transformers (BERT), offering an in-depth analysis of their features and relevance to this research. Section 3 outlines this study's methodology, including data collection and finetuning. In section 4, the authors present the results and engage in a comprehensive discussion, unraveling the significance and implications of their findings. Finally, the paper concludes succinctly, summarizing the key takeaways, and suggesting potential avenues for future research.

2. LITERATURE REVIEW

2.1. Language modeling

Language modeling (LM) is one of the fundamental objectives of natural language processing. LMs attempt to predict the $(n + 1)^{th}$ token in a sequence given the n preceding tokens [4] in order to represent the likelihood of word sequences. Real-life applications of trained LMs are diverse, and various, and include: ClinicalBERT [5], a language model trained on clinical notes to predict hospital readmission, MoIE [6], a molecular foundation language model for drug discovery, that predicts properties based on chemical structure, and sentiment analysis based on LM for general US elections, Biden vs Trump [7]. Four main stages of development/types can be identified for language models:

2.1.1. Statistical language models

They focus on estimating the probability distribution of various linguistic units, such as words, sentences, and entire documents over all possible phrases, in an effort to capture the patterns of natural language [8]. The purpose is to improve the performance of several natural language processing applications. Major Statistical language models (SLM) techniques include but are not limited to: n-grams, decision tree models, linguistically motivated models, exponential models, and adaptive models.

2.1.2. Neural language models

Neural language models (NLMs) anticipate a set of future words given the history of previous words, similar to other language models. However, in NLMs, words are projected onto a lower dimensional vector space via a hidden layer from a sparse, 1-of-V encoding (where V is the size of the vocabulary) [9]. Each word in the vocabulary is depicted as a real-valued feature vector so that, for semantically related words, the cosine of the angles between these vectors is high.

2.1.3. Pre-trained language models

Among all deep learning methods, fine-tuning a pre-trained language model (PLM) on downstream tasks of interest has become the standard pipeline in NLP tasks. The fundamental concept is to use the resulting representations on tasks after training a large generative model on massive corpora, where labeled data is scarce [10]. PLMs are capable of capturing the meaning of words dynamically in consideration of their context. The very first versions of generative pre-trained transformer (GPT) [11] by Open AI, and Google's own BERT [12] are wildly used PLMs that have become very popular in recent years.

2.1.4. Large language models

Two major developments marked the advancement of LLMs: the use of transformer architectures, and the introduction of fine-tuning for pretrained models. A LLM is fine-tuned by further training it on a smaller, task-specific dataset to adapt it to a particular task or topic [13]. The process involves adjusting the parameters of an already trained model using a smaller, domain-specific dataset, and tweaking the weights and parameters of the model to reduce the loss function and enhance its task performance.

LLMs, such as Bard by Google, or ChatGPT by OpenAI, gained exponential popularity over the last couple of years, mainly because the technology matured enough to become viable for consumer use. Although others existed before, it was ChatGPT that made the breakthrough for users outside of the research community and AI enthusiasts, by allowing anyone to sign up for free, open a chat dialogue, and start getting legible LLM output without needing to understand the concept or prompt engineering.

In addition to presenting various opportunities, the utilization of LLMs is accompanied by various ethical and social challenges. Weidinger *et al.* [14] categorized the risks presented as such: i) discrimination, by perpetuating social stereotypes and biases, ii) information hazards, which involves disclosing sensitive information, iii) misinformation hazards, by disseminating misleading information, iv) malicious use, as users with bad intent could take advantage and generate personalized scams and fraud, v) human-computer interaction harms, meaning users may use the humanlike capabilities of LLMs in an unsafe manner, and vi) automation and environmental harms, as training and operating LLMs require extensive computing power, incurring high environmental costs. Some of the main limitations and challenges these LLMs face include the high cost it takes to train them [15], the hallucinations [16] incorrect and/or inappropriate answers given confidently, the bias they might present due to the training data they have been given [17] and their vulnerabilities to adversarial attacks [18].

There are several methods utilized to evaluate the performance of LLMs. Seeing that LLMs can carry out different tasks, it is important to pinpoint what to evaluate, where, and how. These evaluation methods [19] include: measuring features such as language fluency, coherence, context, and more, zero-shot evaluation, which measures the performance of large language models on tasks they have not been explicitly trained on, utilizing other large language models for evaluation, which offers scalability and helps identify areas for improvement, and conceptual evaluation frameworks.

2.2. Large language models in focus

2.2.1. PaLM 2 by Google

When it comes to AI breakthroughs over the years, Google stands out as a household name. For instance, they are the ones behind LLMs all around, as they are based on transformer, a neural network architecture Google made public back in 2017 [20]. PaLM 2 is Google's next-generation large language model that offers a significant improvement from their first one, pathways language model (PaLM), and that was announced in May 2023. It is a 340 billion parameter model trained on 3.6 trillion tokens [21], although much smaller than its predecessor, they found that the training compute surpasses size for the resulting quality.

PaLM 2 is available in four sizes: Gecko, Otter, Bison, and Unicorn, from smallest to largest [22]. Gecko is so small and light that it can run on mobile devices. PaLM 2 can be adapted to accommodate entire product classes in more ways and benefit more users because of its versatility.

PaLM 2 can perform a wide range of tasks [23], all language-related. For instance, it can understand over 100 languages, which allows it to be able to translate and generate text (poems and stories) fluently. PaLM 2 performs exceptionally well in language proficiency exams and passes the advanced levels. It also can logically solve complex problems, whether it is mathematical, riddles, or else. PaLM 2 is the first language model to perform at an expert level with more than 85% accuracy on questions comparable to those on the U.S. Medical Licensing Examination (USMLE). PaLM 2 goes beyond natural languages; it is also capable of coding in over 20 languages, generating code to answer specific problems, debugging it, or even transforming code from one language to another.

Up until now, Google integrated PaLM 2 in nearly 25 of its products, to make them more intelligent. YouTube, Gmail, Google Docs, Google Sheets, and even Google Translate are only some of them. They also recently revealed Google Bard AI, a text-based artificial intelligence chatbot that is built on PaLM 2. Bard generates real-time answers using NLP and machine learning. It takes Google from being a search engine to a capable virtual assistant that provides well-rounded answers from a natural language query [24]. To advance medical LLM efforts, Google's Med-PaLM 2 uses a combination of base LLM advancements, medical domain finetuning, and prompting tactics, including a novel ensemble refinement methodology [25]. This is particularly important as it provides high-quality answers to complex medical questions.

2.2.2. GPT-3.5 by OpenAI

Prior to the development of models like GPT-1, natural language understanding (NLU) tasks relied heavily on supervised learning with labeled datasets, limiting their ability to generalize. GPT-1, a generative pre-trained language model, introduced a transformation by using unsupervised learning on massive unlabeled data, simplifying fine-tuning for downstream tasks [26]. GPT-2, building on GPT-1, improved model structure, increased training data, and scaled up parameters. This enhancement boosted performance across various tasks, reducing the need for supervised training and highlighting the potential of larger models. GPT-3, an even larger model with more data, broke records in language models. It excelled in zero-shot and few-shot settings, demonstrating remarkable versatility across tasks like math, article generation, and coding. As models grow in size, their capabilities continue to expand, promising even more powerful AI systems in the future.

GPT-3.5 is a remarkable milestone in the realm of natural language processing. It represents the evolution of OpenAI's groundbreaking transformer-based models, taking advantage of the lessons learned

from its predecessors, GPT-3 and GPT-2, as well as the foundational transformer architecture. Instead of making GPT-3.5 available in its fully trained state, OpenAI used it to create several systems that are each specially optimized for a different job and are all accessible through the OpenAI API. One of them, text-Davinci-003, is believed to be able to handle more complex commands and generate better-quality, longer-form writing than models built on GPT-3.

In real-life applications, GPT-3.5 has left an indelible mark. Its versatility extends to a wide range of fields and industries. It has been harnessed for natural language understanding tasks, enabling it to extract insights from vast volumes of text data. Content generation, including article writing [27] and creative storytelling, has seen significant improvements thanks to GPT-3.5's ability to produce coherent and contextually relevant text. In the realm of customer service and support, GPT-3.5 has powered chatbots and virtual assistants [28], providing efficient and responsive interactions. Additionally, it plays a pivotal role in automated translation [29], making multilingual communication more accessible.

2.2.2. BERT by Google

BERT which stands for bidirectional encoder representations from transformers was published by researchers at Google AI Language [12] in 2019. It is an innovative language representation model and presents state-of-the-art results in a wide variety of NLP tasks. The main purpose behind BERT is to improve language understanding by pre-training on large amounts of unlabeled data, so it can detect and learn the patterns between the encountered words, terms, and phrases.

The main technical innovation of BERT is the application of transformer's bidirectional training, an extremely popular attention model, to language modeling. In contrast, past studies considered text sequences from either a left-to-right or a combined left-to-right and right-to-left training perspective [30]. The results of the study show that bidirectionally trained language models have a better ability to comprehend context and language flow more deeply than single-direction language models. Although it was previously not practicable, bidirectional training became possible because of masked LM (MLM) on which BERT relies. BERT is versatile, and the tasks it can perform are numerous. Among its applications, notable examples include: text classification [31], question answering [32], natural language inference [33], pre-training language representations [12], sentiment analysis [34], fake news detection [35], and more.

For many use cases, BERT proved to perform better than many models. However, it also presents several limitations, such as the limit on the size of input text that it can handle, which is a challenge when dealing with long documents or summaries of hefty articles [36]. Its vocabulary is fixed, which prevents it from handling rare or out-of-vocabulary terms. Although it can be fine-tuned for specific tasks, its performance might be below standard if they are significantly different from the ones it was trained on [37]. Another limitation of BERT is its high computational expense, especially when it comes to larger models.

BERT cannot be discussed without addressing the transformer architecture, which is a pivotal advancement in natural language processing (NLP) that revolutionized the way language models are designed and trained. It was introduced in the paper titled "Attention is All You Need" by Vaswani *et al.* [20] and quickly became the foundation for many state-of-the-art language models. Its key components include: i) a self-attention mechanism, which enables the model to understand dependencies between the different input text parts, by calculating attention weights for each token; ii) multi-head attention, and every single head learns different relationships between the words, capturing different aspects of the context, positional coding; iii) positional encodings, that are added to the word embeddings to provide information about the word's position in the sequence; iv) feedforward neural networks, enabling the model to learn complex interactions and transformations; and v) layer normalization and residual connections, which are techniques used to stabilize training and enable the network to learn effectively in very deep architectures.

Some of the benefits of the transformer architecture for LLMs are: parallelism, meaning the self-attention mechanism allows the model to process different parts of the input sequence in parallel, which significantly speeds up training and inference compared to RNNs or CNNs, its ability to capture long-range dependencies, making it ideal for understanding a word within an entire document, generating more coherent and contextually appropriate text because it captures contextual information for both left and right of a word, and its scalability, so it handles larger datasets and more complex tasks.

3. METHOD

In their endeavor to assess the performance of the three prominent LLMs: GPT-3.5, PaLM 2, and BERT, in digital marketing, the authors embarked on a process of fine-tuning these models. Their primary objective centered around the generation of compelling email subjects tailored for promotional campaigns. Not only that, but the email subjects needed to land in the primary inbox, avoiding the spam folder at all

costs. This involved refining the models using a dataset that encompassed a wide range of subject lines, that were previously tested out, allowing the models to better comprehend the language specific to the email marketing domain. Through this experiment, the aim is to dissect and contrast the effectiveness of each LLM's proposed email subjects by evaluating the percentage of recipients who engage by opening these emails. This comprehensive analysis promises insights into the diverse capabilities of these LLMs in crafting subject lines that resonate with recipients and drive higher email open rates.

For this paper, the authors have partnered with an undisclosed email marketing company, which has granted us access to a current database comprising 644,600 email subjects. This dataset is structured as such: the advertised product name, and the email subject that's been used in the campaign, accompanied by corresponding statistics, comprising the number of deliveries and openings associated with each email subject, which gives us the opening rate. The percentage of openings is the main measure of the impact of an email subject, as it is the first thing the customer sees. Due to the company's preference for anonymity, their identity remains undisclosed.

Beginning with the aforementioned CSV dataset containing important information including product names, corresponding email subjects, as well as delivery and open statistics, the authors initiated the customary phase of data cleaning. This involved actions such as eliminating blanks, and addressing inconsistencies to ensure the dataset's integrity. Subsequently, they executed a strategic transformation, converting each dataset entry into a pair of specialized prompt and output structure. These prompts were meticulously designed to solicit email subject suggestions for a given product, aiming to achieve a specified target opening rate. These prompts were coupled with their corresponding answers, the original email subjects. LangChain was used for this.

The objective behind LangChain is to prepare data to be accessed by the LLM with the least amount of computation possible. It does that by breaking down data into smaller "chunks" that get embedded into a vector store [38]. Moving forward, when it is time to create the prompt-completion pair, the vectorized representations of the large document can be used in addition to the LLM. LangChain also facilitates other API-related tasks, such as surfing the web, sending emails and creating prompts. Finetuning LLMs involves re-training pre-trained models on extensive datasets, which will customize the model to be efficient at the specific task at hand [39]. This technique has revolutionized natural language processing (NLP) by adapting pre-trained models to specific applications, such as question-answering, language translation, named-entity recognition, document summarization, sentiment analysis, and more [40].

For the fine-tuning process of PaLM 2, the authors chose the text-bison@001 model as its objectives are the closest to theirs, and they strategically ran 3 epochs, meticulously chosen to align with the dataset's characteristics and the batch size and steps. By utilizing the Adam optimizer with a default learning rate of 0.001, the authors ensured optimal convergence during training. Additionally, their implementation integrated TensorBoard, enabling thorough monitoring and analysis of training metrics, thus facilitating comprehensive performance evaluation post-training.

To ensure an accurate and fair comparison, the authors extended their fine-tuning efforts to include the GPT-3.5 and Bert models, employing similar hyperparameters as those utilized for PaLM 2, the same optimizer, learning rate and the number of steps. This approach guarantees that the comparison between PaLM 2, GPT-3.5, and Bert is conducted under consistent conditions, allowing for a meaningful assessment of their respective performance and effectiveness in addressing the authors' objectives.

To comprehensively evaluate the performance of the finetuned LLMs, BLEU and ROUGE scores were used to compare. Although the authors have the model accuracy in hand, using it for text generation, as opposed to classification or other, is meaningless. That is because text generation allows for creative freedom, unlike other tasks where the answer is either right or wrong. BLEU and ROUGE-LSum are widely employed evaluation metrics in natural language processing, notably for machine translation and summarization. BLEU assesses the extent of overlap between a reference sentence (X) and a candidate sentence (Y) [41], while ROUGE-LSum evaluates the longest common subsequence of tokens between a candidate (Y) and reference (X), specifically designed for assessing summary quality.

Afterward, to see the performance of the finetuned models in real life, the authors opted to test 3 products that had undergone previous human-designed campaigns and achieved different opening rates, and generated email subjects to promote each using the three finetuned LLMs: GPT-3.5, PaLM 2, and BERT. These subjects were then distributed to an equivalent number of email addresses. While the addresses were not identical, they did share a demographic similarity, ensuring a consistent context for comparison. Additionally, all the selected addresses were of excellent quality, marked by their active engagement and recent interactions. This meticulous approach allowed the authors to gauge the LLMs' efficacy in generating subject lines that rival both human efforts and each other. To sum it up, these models were evaluated by several methods: using the customary BLEU and ROUGE scores, their training accuracy, and finally testing their generated subject lines in real life, calculating their open rates.

4. RESULTS AND DISCUSSION

Results were meticulously gathered over 3 days following the campaign launches, ensuring a robust dataset for analysis. The chosen timeframe allowed for a comprehensive assessment of recipient engagement while minimizing the impact of daily fluctuations. Rigorous validation procedures were implemented to maintain the accuracy of the final findings. These included validating the email list by cleaning it from inactive or invalid addresses, implementing permission-based opt-in, handling unsubscribing, and tracing the actions on the email, such as opening, sending to spam and clicking on the link.

Table 1 presents three performance metrics for the finetuned language models: PaLM 2, GPT-3.5, and BERT, represented as columns, while each row represents a different evaluation metric: BLEU score, ROUGE-LSum score, and training accuracy. The evaluation results show that both BERT and PaLM 2 outperform GPT-3.5. A higher BLEU score indicates better precision, while a higher ROUGE-LSum score signifies improved recall. Therefore, the finetuned GPT-3.5 model exhibits the lowest precision, recall, and accuracy among the three models when compared against human-created content. The performance variations across these metrics are attributed to each model's unique capabilities. BERT's bidirectional context comprehension enables it to excel in tasks like email subject generation, resulting in superior BLEU and ROUGE-LSum scores. PaLM 2, optimized for conversational interactions, achieves slightly lower performance levels. On the other hand, despite its extensive knowledge base, GPT-3.5 may lack fine-tuning for this specific task, leading to overall lower scores.

Table 1. Finetuned LLMs BLEU and ROUGE-LSum scores and training accuracy

	PaLM 2	GPT-3.5	BERT
BLEU score	0.59	0.32	0.62
ROUGE-LSum score	0.72	0.58	0.76
Training accuracy	0.66	0.51	0.68

The email open rate, a key metric in marketing, measures the percentage of recipients who open commercial emails they receive. It is calculated by dividing the number of opened emails by the total emails delivered. Factors influencing open rates include industry type, timing of delivery, and email subject lines. To evaluate the impact of subject line quality, the authors maintained consistent parameters and compared open rates between emails with subject lines generated by the finetuned models against those crafted by humans.

Table 2 presents the opening rates of promotional emails for three different products, of which the subjects are generated by the finetuned LLMs. The results are as follows: all three tested LLMs, namely GPT-3.5, BERT, and PaLM 2, demonstrated superior performance when compared to human-generated efforts. However, a closer examination of the opening rates reveals that BERT and GPT-3.5 exhibited remarkably similar results, showcasing their comparable effectiveness in crafting engaging subject lines. In contrast, PaLM 2, while still outperforming human-generated subjects, displayed slightly lower opening rates, indicating a modest gap in its ability to capture recipient attention compared to its AI counterparts. These findings underscore the advancements in AI-driven text generation and their potential to surpass human-generated content in optimizing email marketing strategies. These results go hand in hand with the BLUE and ROUGE scores each finetuned model had, as the finetuned BERT model emulates human-like responses the best, replicating the deficiencies as well. Considering that the human-generated content itself has performed the worst compared to others, this supports the obtained results.

All three of these LLMs are impressive in terms of tasks they can perform, and the performance they presented. However, each of them has its own strengths, which led to the discrepancy in the results. Due to its bidirectional nature, BERT is known to be more advantageous in terms of natural language understanding (NLU) and sentiment analysis, as opposed to GPT and PaLM 2 which perform better at common sense tasks, pragmatic inference, and text generation. An important pillar of digital marketing is understanding the customers' point of view and showing them engaging content that will lead to conversions eventually, which could explain why BERT was better.

Also, the variations in opening rates could be a result of different optimization techniques employed for each type of subject line. While AI models like PaLM 2, GPT-3.5, and BERT can undergo iterative A/B testing at a larger scale and with more data, humans may have limitations in terms of conducting extensive experiments. AI models can fine-tune subject lines based on massive data they analyze and learn the patterns of what works and what does not. Human-generated subject lines may not undergo the same level of rigorous optimization, leading to lower opening rates.

These LLMs have one major thing in common: they all have the transformer architecture as a foundation. This architecture revolutionized the NLP field and appears to be going nowhere, as most recent LLMs are based on it and are drifting away from sequential processing models, such as RNNs and LSTMs for instance. That is because transformer's key innovation, attention mechanisms, improves the model's ability to handle long-range dependencies within text, which is ideal for text generation, among other tasks.

It appears from the results of the current study that incorporating LLMs into email marketing campaigns has produced positive outcomes. However, the evolving nature of technology and consumer behavior points to several promising directions for further study and real-world application. First, continued improvements in LLMs, such as the development of more complex and domain-specific models, may improve their effectiveness in creating intriguing email subject lines. Moreover, as privacy regulations evolve and consumer preferences change, investigating the ethical considerations and personalization challenges associated with AI-generated content remains imperative. Furthermore, examining the long-term effects of AI-generated email subjects on customer engagement and brand loyalty will provide a comprehensive understanding of their role in shaping future marketing strategies, and may be applied to other components: email body, titles of ads and article titles.

Table 2. Emails open rates with LLMs and human generated subjects

Opening rate	PaLM 2	GPT-3.5	BERT	Human Generated
Product 1	39.5%	44.8%	47.7%	28.4%
Product 2	14.6%	21.7%	20.3%	6.8%
Product 3	28.7%	31.4%	33.8%	9.2%

5. CONCLUSION





In summary, this paper explored the efficacy of email subject lines generated by LLMs in the context of digital marketing, more specifically email marketing campaigns. The primary objectives of this study were to assess the performance of LLMs, namely GPT-3.5, BERT, and PaLM 2, in comparison to human-generated subject lines, and to discern the factors contributing to variations in opening rates across different products. The findings have revealed that all three LLMs surpassed human-generated content in terms of capturing recipient attention, signifying the potential of AI-driven text generation in email marketing. Moreover, while GPT-3.5 and BERT demonstrated similar performance, PaLM 2 exhibited a slightly lower open rate, suggesting room for improvement. The discussion encompassed factors such as semantic coherence, psychological triggers, and optimization strategies, shedding light on the mechanisms behind the observed results. Looking ahead, the dynamic landscape of AI and email marketing presents exciting avenues for future research, including the development of domain-specific LLMs, the exploration of hybrid AI-human approaches, and the examination of ethical considerations. As this evolving field is opening to more and more opportunities, this study contributes valuable insights to guide marketers and researchers toward a deeper understanding of AI's role in shaping the future of email marketing strategies.

REFERENCES





- [1] P. van Esch and J. Stewart Black, "Artificial intelligence (AI): revolutionizing digital marketing," *Australasian Marketing Journal*, vol. 29, no. 3, pp. 199–203, Aug. 2021, doi: 10.1177/18393349211037684.
- [2] A. K. Kushwaha and A. K. Kar, "Language model-driven chatbot for business to address marketing and selection of products," in *IFIP Advances in Information and Communication Technology*, vol. 617, Springer International Publishing, 2020, pp. 16–28.
- [3] M. Firat, "How chat GPT can transform autodidactic experiences and open education? future of education view project visual perception in educational design view project," Jan. 2023.
- [4] S. Merity, N. S. Keskar, and R. Socher, "An analysis of neural language modeling at multiple scales," *arXiv preprint arXiv:1803.08240*, Mar. 2018.
- [5] O. J. B. D. Walk Iv, T. Sun, A. Perotte, and N. Elhadad, "Clinically relevant pretraining is all you need," *Journal of the American Medical Informatics Association*, vol. 28, no. 9, pp. 1970–1976, Jun. 2021, doi: 10.1093/jamia/ocab086.
- [6] O. Méndez-Lucio, C. Nicolaou, and B. Earnshaw, "MolE: a molecular foundation model for drug discovery," *arXiv preprint arXiv:2211.02657*, Nov. 2022.
- [7] R. Chandra and R. Saini, "Biden vs trump: modeling US general elections using BERT language model," *IEEE Access*, vol. 9, pp. 128494–128505, 2021, doi: 10.1109/ACCESS.2021.3111035.
- [8] R. Rosenfeld, "Two decdes of statistical language modeling where do we go form here? Where do we go from here?," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1270–1275, Aug. 2000, doi: 10.1109/5.880083.
- [9] Y. Kim, Y. I. Chiu, K. Hanaki, D. Hegde, and S. Petrov, "Temporal analysis of language through neural language models," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 61–65, May 2014, doi: 10.3115/v1/w14-2517.
- [10] S. Edunov, A. Baevski, and M. Auli, "Pre-trained language model representations for language generation," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, Mar. 2019, vol. 1, pp. 4052–4059, doi: 10.18653/v1/n19-1409.

- [11] G. Chalvatzaki, A. Younes, D. Nandha, A. T. Le, L. F. R. Ribeiro, and I. Gurevych, "Learning to reason over scene graphs: a case study of finetuning GPT-2 into a robot language model for grounded task planning," *Frontiers in Robotics and AI*, vol. 10, Aug. 2023, doi: 10.3389/frobt.2023.1221739.
- [12] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, Oct. 2019.
- [13] R. Behnia, M. R. Ebrahimi, J. Pacheco, and B. Padmanabhan, "EW-Tune: a framework for privately fine-tuning large language models with differential privacy," in *IEEE International Conference on Data Mining Workshops, ICDMW*, Nov. 2022, vol. 2022-Novem, pp. 560–566, doi: 10.1109/ICDMW58026.2022.00078.
- [14] L. Weidinger *et al.*, "Taxonomy of risks posed by language models," in *ACM International Conference Proceeding Series*, Jun. 2022, pp. 214–229, doi: 10.1145/3531146.3533088.
- [15] O. Sharir, B. Peleg, and Y. Shoham, "The cost of training NLP models: a concise overview," *arXiv preprint arXiv:2004.08900*, Apr. 2020.
- [16] N. McKenna, T. Li, L. Cheng, M. J. Hosseini, M. Johnson, and M. Steedman, "Sources of hallucination by large language models on inference tasks," *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2758–2774, May 2023, doi: 10.18653/v1/2023.findings-emnlp.182.
- [17] E. Ferrara, "Should ChatGPT be biased? challenges and risks of bias in large language models," *First Monday*, vol. 28, no. 11, Apr. 2023, doi: 10.5210/fm.v28i11.13346.
- [18] R. Pedro, D. Castro, P. Carreira, and N. Santos, "From prompt injections to SQL injection attacks: how protected is your LLM-integrated web application?," *arXiv preprint arXiv:2308.01990*, Aug. 2023.
- [19] Y. Chang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, Mar. 2024, doi: 10.1145/3641289.
- [20] A. Vaswani *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 5999–6009, Jun. 2017.
- [21] J. Elias, "Google's PaLM 2 uses nearly five times more text data than predecessor," <https://www.cnn.com/2023/05/16/googles-PaLM-2-uses-nearly-five-times-more-text-data-than-predecessor.html> (accessed May 16, 2023).
- [22] Google, "Google AI PaLM 2," Google AI, 2023. <https://ai.google/discover/palm2/> (accessed May 16, 2023).
- [23] R. Anil *et al.*, "PaLM 2 technical report," *arXiv preprint arXiv:2305.10403*, May 2023.
- [24] Ö. Aydın, "Google bard generated literature review: metaverse," *Journal of AI*, vol. 7, no. 1, pp. 1–14, Dec. 2023, doi: 10.61969/jai.1311271.
- [25] K. Singhal *et al.*, "Towards expert-level medical question answering with large language models," *arXiv preprint arXiv:2305.09617*, May 2023.
- [26] M. Zhang and J. Li, "A commentary of GPT-3 in MIT technology review 2021," *Fundamental Research*, vol. 1, no. 6, pp. 831–833, Nov. 2021, doi: 10.1016/j.fmre.2021.11.011.
- [27] O. Buruk, "Academic writing with GPT-3.5 (ChatGPT): reflections on practices, efficacy and transparency," *ACM International Conference Proceeding Series*, pp. 144–153, Feb. 2023, doi: 10.1145/3616961.3616992.
- [28] A. Shafeeg, I. Shazhaev, D. Mihaylov, A. Tularov, and I. Shazhaev, "Voice assistant integrated with chat GPT," *Indonesian Journal of Computer Science*, vol. 12, no. 1, Feb. 2023, doi: 10.33022/ijcs.v12i1.3146.
- [29] W. Jiao, W. Wang, J. Huang, X. Wang, S. Shi, and Z. Tu, "Is ChatGPT a good translator? yes with GPT-4 as the engine," *arXiv preprint arXiv:2301.08745*, Jan. 2023.
- [30] L. Zhou, J. Zhang, and C. Zong, "Synchronous bidirectional neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 91–105, Apr. 2019, doi: 10.1162/tacl_a_00256.
- [31] Q. Yu, Z. Wang, and K. Jiang, "Research on text classification based on BERT-BiGRU model," *Journal of Physics: Conference Series*, vol. 1746, no. 1, Jan. 2021, doi: 10.1088/1742-6596/1746/1/012019.
- [32] Z. Wang, P. Ng, X. Ma, R. Nallapati, and B. Xiang, "Multi-passage BERT: a globally normalized BERT model for open-domain question answering," *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 5878–5882, Aug. 2019, doi: 10.18653/v1/d19-1599.
- [33] N. Jiang and M. C. de Marneffe, "Evaluating BERT for natural language inference: a case study on the CommitmentBank," in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2019, pp. 6086–6091, doi: 10.18653/v1/d19-1630.
- [34] X. Li, L. Bing, W. Zhang, and W. Lam, "Exploiting bert for end-to-end aspect-based sentiment analysis," in *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pp. 34–41, Oct. 2019, doi: 10.18653/v1/d19-5505.
- [35] S. Kula and R. C. Michałand Kozik, "Application of the bert-based architecture in fake news detection," in *Advances in Intelligent Systems and Computing*, vol. 1267 AISC, Springer International Publishing, 2021, pp. 239–249.
- [36] M. Khadhraoui, H. Bellaaj, M. Ben Ammar, H. Hamam, and M. Jmaiel, "Survey of BERT-base models for scientific text classification: COVID-19 case study," *Applied Sciences (Switzerland)*, vol. 12, no. 6, Mar. 2022, doi: 10.3390/app12062891.
- [37] Y. Lu, J. Pan, and Y. Xu, "Public sentiment on Chinese social media during the emergence of COVID19," *Journal of Quantitative Description: Digital Media*, vol. 1, Apr. 2021, doi: 10.51685/jqd.2021.013.
- [38] Langchain, "Introduction langchain," *Python.Langchain.Com*, 2023. <https://python.langchain.com/docs/community> (accessed May 16, 2023).
- [39] S. Ruder, M. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing tutorial," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Tutorial Abstracts*, 2019, pp. 15–18, doi: 10.18653/v1/n19-5004.
- [40] A. Madaan, S. Zhou, U. Alon, Y. Yang, and G. Neubig, "Language models of code are few-shot commonsense learners," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, 2022, pp. 1384–1403, doi: 10.18653/v1/2022.emnlp-main.90.
- [41] S. Loukili, A. Fennan, and L. Elaachak, "Applications of text generation in digital marketing: a review," *ACM International Conference Proceeding Series*, May 2023, doi: 10.1145/3607720.3608451.





BIOGRAPHIES OF AUTHORS

Soumaya Loukili     is a dedicated Ph.D. student, passionate about the synergy between marketing, artificial neural networks, and artificial intelligence. Her research explores innovative approaches in these fields, showcasing a commitment to staying at the forefront of cutting-edge research. Soumaya's work reflects a fresh perspective on marketing, leveraging the power of artificial neural networks and AI for insightful strategies. She also has a keen interest in sentiment analysis and did multiple studies in this field, intersecting with artificial intelligence as well. She can be contacted via email at: soumaya.loukili1@etu.uae.ac.ma.



Abdelhadi Fennan     is a professor in the Department of Computer Science at Abdelmalek Essaadi University, Tangier, Morocco. Holding a validated email address at uae.ac.ma, he is a distinguished academic with a focus on artificial intelligence, software engineering, and information systems (business informatics). His expertise extends to disciplines such as higher education, educational technology, and curriculum theory. Abdelhadi's skills encompass e-learning, online learning, semantic web, sentiment analysis, and technology-enhanced learning. With a strong commitment to research and academia, he plays a vital role in advancing knowledge and innovation in these domains. He can be contacted via email at: afennan@gmail.com.



Lotfi Elaachak     is an associate professor at Faculté des Sciences et Techniques de Tanger. With a strong focus on artificial intelligence, he has a profound understanding of algorithms and artificial neural networks. His expertise extends to higher education, where he actively engages in innovative teaching methods. Proficient in machine learning, Lotfi has made significant contributions to the field, particularly in the realms of neural networks and artificial intelligence. Beyond academia, he is involved in the realm of serious games, video games, game design, and game development, showcasing a diverse skill set. Lotfi is passionate about incorporating these skills into teaching methods, emphasizing the integration of digital games, game-based learning, and new technologies like virtual reality and augmented reality. His proficiency also extends to Arabic natural language processing (NLP), reflecting his commitment to advancing research and education in these dynamic domains. He can be contacted via email at lelaachak@uae.ac.ma.