Generative AI for Consumer Communications: Classification, Summarization, Response Generation

Nelson Correa
Andinum, Inc.
West Palm Beach, FL, USA
ncorrea@ieee.org

Antonio Correa
Andinum, Inc.
West Palm Beach, FL, USA
antonio@andinum.com

Wlodek Zadrozny
UNC Charlotte, Computer Science and Data Science
Charlotte, NC, USA
wzadrozn@charlotte.edu

***Abstract*—Generative AI showed the unexpected power of large language models (LLMs) for understanding and generation of natural language text and other modalities at the end of 2022. This paper presents a novel generative AI system for text classification, summarization and response generation of consumer communications. The system uses the same foundation model and a uniform pipeline for the tasks proposed. Consumer communications are massive and served mainly via voice and text, and until recently could be handled only with human agents (customer service representatives). However, they must be handled with quality, consistency, speed and low cost, at scale. We limit our attention to financial consumer communications from the U.S. Consumer Financial Protection Bureau (CFPB), publicly available in a dataset of over 4.7 million complaints. Performance reaches 88% accuracy (*without fine-tuning*) for classification and over 72% for summarization and response generation. Artificial intelligence has great positive impacts for business and society, but its application and deployment also poses risks and unknowns. We thus address the important questions of risk, bias, interpretability, explainability, safety and regulatory compliance with the emerging legal frameworks.**

***Index Terms*—Generative AI; natural language processing; large language models; text classification; text summarization; vector embeddings; AI safety**

I. Introduction

A. A few words about Generative AI

The encoder-decoder transformer architecture [1] and the neural attention mechanism underlying it [2] first appeared in 2017 and 2015, respectively. BERT, an encoder-only transformer implementation from Google, with 340 million parameters for BERT-large, was released soon after, in 2018 [3]. Similarly, the Generalized Pre-trained Transformer (GPT) of Alec Radford and team at OpenAI was released in 2018 [4]. GPT-2, GPT-3 and GPT-4, new versions of GPT, rapidly followed, with the number of parameters increasing from 117 million in the original GPT, to 1.5 billion in GPT-2, 175 billion in GPT-3, and an estimated 1.75 trillion in the most recent GPT-4 [5], [6], and the immensely popular ChatGPT [7].

These new large language models, with billions of parameters programmed by machine learning from large amounts of data, can now accomplish a few tasks of general artificial intelligence, at a level that matches or surpasses human performance in some tasks. The tasks can broadly be classified as predictive (e.g., regression and classification) or generative (e.g., text or image generation). Evaluation of the later, however, now broadly called "*Generative AI*," is more elusive.

The unexpected power of large language models for understanding and generation of text and other modalities became apparent in November, 2022, with the release of OpenAI's ChatGPT (GPT-3.5). In parallel with the release of the GPT models, there has been rapid progress and release of alternative commercial and open source LLM offerings from Google (LaMDA, BARD and Gemini [8]), the creator of the original transformer architecture; Meta (LLaMA [9]); Anthropic (Claude); HuggingFace and others. A Feb'24 article provides a comprehensive list and overview of their capabilities [10].

On the applications side, businesses and organizations must regularly handle large volumes of "unstructured data" in the form of text documents, voice, images, sensor data and video. In a large global enterprise, the volume is of billions of documents per year (e.g., client and employee communications), that until recently could be handled only with large scale human input (call centers, line-of-business and customer service representatives), with significant IT application support. Understanding unstructured data and acting on it requires perceptual abilities and strategic decision making that until today requires human cognitive ability, even for clerical and routine tasks, like communicating with a client. The gap is, however, closing. Two recent books and a collection of popular articles [11]–[13] discuss the challenges and promises of AI.

B. Focus and contributions of this article

In this paper we present a novel generative AI application that accomplishes three document processing tasks: classification, summarization and response generation for textual consumer communications. The system uses the same underlying foundation large language model and a uniform pipeline for accomplishing the three tasks mentioned.

We use the recent Google Gemini Pro large language model and API [8], [14] (March 2024). In addition we use standard Python libraries for numerical computation, data science, machine learning, visualization, networking, storage and application development.

We work here with two subsets of consumer complaints from a publicly available dataset of over 4.7 million complaints, from the U.S. Consumer Financial Protection Bureau (CFPB), corresponding to years 2011-2016 [15] and 2023 [16].

979-8-3503-5528-4/24/$31.00 ©2024 IEEE

The contributions of this paper are:

- Presentation of large language models and generative AI from research, technical and application perspectives;
- Description and use of a real world large dataset (CFPB) for development and evaluation of the new system;
- Development of a novel generative AI system for classification, summarization and response generation with high performance (88% accuracy for classification; over 72% for summarization);
- Comparison of classification accuracy on one of the two CFPB datasets using, among others, two sets of embeddings, FinBERT [17] and Gemini [8], showing that a fine-tuned smaller large language model can perform equally well (confirming an observation in [18]).
- Use of the emerging methodology and issues relevant to the development and deployment of new AI applications, including model bias, interpretability, explainability, AI safety and recent regulation [19].

The paper is organized as follows: Section II presents prior work and methodology; Section III the CFPB consumer complaints dataset; Section IV machine learning models and tasks; Section V results and comparison; Section VI interpretability, explainability and safety; and Section VII future work.

A GitHub repository with our models, python code, datasets, and Jupyter notebooks will be available for the conference.[1]

II. Prior work and methodology

Artificial intelligence and mathematical analysis of language can be traced to 1950, with the publications by Turing on the concept of intelligence [20], and by Shannon on information theory and the statistical analysis of English [21]. Similarly, methods for the computational analysis of language and speech, and the development of practical applications started in the 1950s using linguistic and symbolic methods [22]–[24], and subsequently using statistical (e.g., for speech recognition and translation) [25] and neural methods [26], [27].

The three topics of interest in this article are text classification, summarization and response generation. The best understood of these is classification, perhaps because its metrics such as precision and recall agree with human intuitions. [28] provides an elementary practical introduction to neural models and classification, including transformers.

Automated summarization is more complicated, and human evaluation does not necessarily agree with automated metrics [29]. For example, despite their potential for hallucination [30], or factual inconsistency [31], summarization via large language models seems to be preferred to other methods [32], such as extractive summarization, where the summary is a subset of the original document.

Finally, since the problem of responding to customers concerns is ubiquitous, and LLMs are potentially useful, the question arises as to how to realize this potential. At this point there is no solution addressing appropriateness of responses (e.g. contradictory responses, changing topics, hallucinations). Three approaches are currently used, prompt engineering [33], LLM system architecture, and fine-tuning [10], with the latter two being perhaps more practical [34] from the point of view of efficacy and cost (confirmed by this paper).

[1]https://nelscorrea.github.io/andescon2024/.

III. CFPB Consumer Complaints Datasets

The CFPB consumer complaints dataset has been collected since the end of 2011 and is receiving approximately one million complaints per year since 2021. As of March, 2024, the dataset comprises over 4,700,000 complaints [16] .

In this paper we access two subsets of the CFPB dataset, with data collected from 2011-2016 and released by Kaggle [15], as well all complaints received in 2023. The Kaggle subset contains 555,957 records; the 2023 subset contains almost 1.3 million records.

However, not all records contain text, and we focus on the records with text. The Kaggle subset has 66,806 records with non-empty complaint narratives, with 12,736,164 total words, over a vocabulary of 49,451 unique words, and a mean document length of 190.6 words per text (minimum of 2 and maximum of 1,284 words). More details can be found in [35]. The 2023 CFPB data set has 486,880 records with text. [2]

A. Data fields

Each record contains 18 fields, with key fields "complaint_id", "date_received", "complaint_what_happend", "company", "state" and "zip_code", "product", "sub_product", "issue" and "sub_issue". The consumer complaint narrative is "complaint_what_happend", and the fields "product" and "issue" can be target classification labels.

For the tasks of interest in this paper, classification, summarization and complaint response generation, we limit ourselves to the fields "complaint_id", "date_received", "company", "complaint_what_happend" and "product."

B. Example consumer complaint narratives

The following are two redacted (*by CFPB*) and shortened (*by us*) extracts of consumer complaint narratives:

complaint_id: 1655441 (150 words) —– "My checking account was charged XXXX overdraft fees over a period of time of approx one week because of an overdraft that then cascaded into fee after fee after fee because of negative balance transactions in the account. [. . .] The bank also only sent me an email alert today XXXX XXXX, of an overdrawn account ... while for days I have been racking up fees. And never was there an alert that I was getting continual fees."

complaint_id: 1290606 (413 words) —– "[Company] used deceptive collection practices while attempting to collect on a purchased debt. Once initial contact had been made and a discussion on solutions to resolve the debt, the attorney office sent out a summons for an appearance in court. [. . .] [Company] is using the power and intimidation of the summons and the court system as a scare tactic and collection tool to coerce and discriminate against consumers."

[2]We haven't measured the vocabulary distribution, but we expect the statistics to be comparable to the Kaggle dataset.

IV. Machine Learning Models and Tasks

We present details of the NLP models for our tasks: Classification, summarization and response generation.

A. Using Google Gemini for dense vector embeddings

A preliminary to all tasks is to create vector representations (*embeddings*) of the input texts (cf. [28]). The Google Gemini API is a variant of the Gemini Ultra pre-trained model, and provides access to models for tasks 'embedContent', and 'generateContent', among others. We use the 'embedding-001' model to compute dense vector embeddings, with a task type set to one of five task types:

TABLE I
Gemini Embeddings Task Types

Task Type	***Description***
RETRIEVAL_QUERY	Specifies the given text is a query in a search/retrieval setting.
RETRIEVAL_DOCUMENT	Specifies the given text is a document in a search/retrieval setting.
SEMANTIC_SIMILARITY	Specifies the given text will be used for Semantic Textual Similarity (STS).
CLASSIFICATION	Specifies that the embeddings will be used for classification.
CLUSTERING	Specifies that the embeddings will be used for clustering.

For the classification task, only the complaint narrative text of the Kaggle dataset was converted to embedding vectors using the 'classification' task type hyper-parameter. For other experiments, both the Kaggle and the 2023 CFPB datasets were converted into embedding vectors by first creating a JSON representation of each record, and then similarly presenting the JSON record as a string to Gemini, in batches of 20K to 40K records using Google Colab Pro (high RAM, no GPU) to create the embedding. The process took approximately 20min/40K records (but required several attempts).

B. Data visualization

Data visualization helps to understand complex datasets. Fig. 1 shows a cluster visualization of the CFPB Kaggle dataset by 'product' category, using Google Gemini vector embeddings of the dataset and a 2-D dimensionality reduction with the t-SNE technique.

C. Classification with dense embeddings and head MLP

Our complaint classification model (CFPB 'Product' label) uses the Google Gemini 1.0 API, model 'embedding-001' to produce a 768-dimensional embeddings for the input complaints, followed by a multi-layer perceptron (MLP) with one hidden layer of the same dimension as the input, a 'dropout' layer with 0.5 dropout, and a 'softmax' output layer with 11 outputs, corresponding to the 11 output label classes.

The MLP model has 599,051 trainable parameters, and is used here as a classification head since this is an efficient, performant and understandable model, that matches the final dense classifier layers of the transformer, as well as the final classifier of previous work [35], enabling better comparison below with prior results.

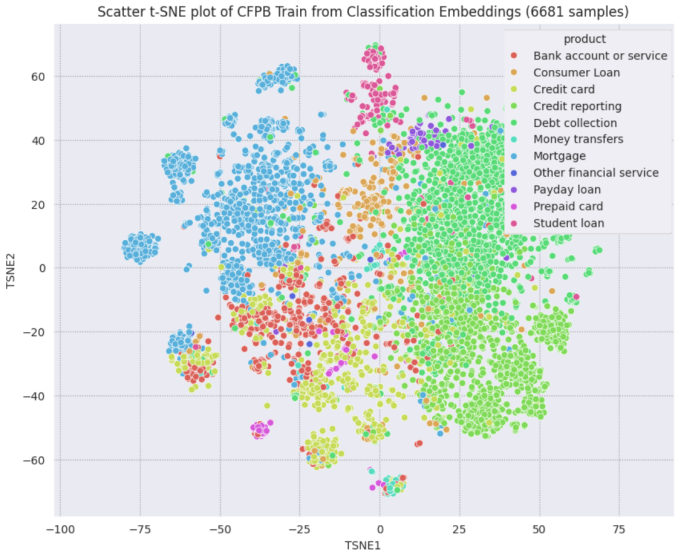

Fig. 1. CFPB Complaint Clusters by 'Product' Category

The CFPB dataset is given an 80/10/10 training/development/test data split, for a training dataset size of 53,444 complaints and development and test set sizes of 6,681 complaints. The MLP is trained for 25 epochs (the performance stabilizes at about 22 epochs), with input batch size of 32. Training takes 4:20 minutes on an Intel i7 8-core processor with 16 GB RAM.

D. Complaint summarization with Google Gemini

The complaint summarization and response generation tasks use the Gemini Pro 'generateContent' model of Google Gemini 1.0 API to generate text outputs.

A basic summarization prompt to the LLM follows:

> Your task is to summarize the following complaint in {num_sents} sentences. Use a maximum {num_words} words for the summary. Summarize the complaint succinctly for a customer service representative.
> COMPLAINT: {complaint}
> SUMMARY:

Two sample summaries for complaint '1290606' above, using our summarization prompt, follow:

complaint_id: 1290606 (413 words)
SUMMARY 1 (60 words, 392 chars):

> [company] allegedly used deceptive collection practices by issuing a court summons while negotiating a payment arrangement. The consumer, an advocate for her mother-in-law, claims she was hung up on multiple times and denied verification of the debt. The summons allegedly arrived just days after initial contact, despite prior assurances that no legal action had been filed.

SUMMARY 2 (48 words, 301 chars):

> [company] sent a summons for court appearance while the consumer was in good faith discussions to resolve a debt, using intimidation tactics to coerce payment. Despite numerous attempts to contact management, the consumer was repeatedly hung up on and denied verification of the debt.

Note the variability of summaries; LLM output is stochastic, and thus the summarization responses can vary for each run. More elaborate summarization prompts are possible, e.g., to include more specific details in the summary.

E. Complaint response generation with Google Gemini

A basic complaint response generation prompt follows:

> Your task is to write an email response to the consumer for the following complaint in {num_sents} sentences. Use a maximum {num_words} words for the response.
> COMPLAINT: {complaint}
> EMAIL RESPONSE:

The following is a sample LLM response to the same complaint '1290606' using this prompt:

EMAIL RESPONSE (39 words, 225 chars):

> Thank you for reaching out to us about this matter. We are investigating your complaint and will be in touch with you soon to discuss it further.
> We appreciate your patience and understanding while we look into this matter.

More elaborate response generation prompts are possible, with more specific details to include in the response. Complaint responses can also show the same variability as before.

V. Results

The evaluation methodology and results for our three tasks with the Gemini large language model follow.

A. Task 1: Complaint classification performance

Fig. 2 shows the model classification report on test data, with precision, recall and F-1 accuracy measure for each of the output classification categories.

	precision	recall	f1-score	support
Bank account or service	0.83	0.85	0.84	571
Consumer Loan	0.82	0.72	0.77	368
Credit card	0.87	0.84	0.86	793
Credit reporting	0.91	0.90	0.90	1253
Debt collection	0.86	0.89	0.88	1756
Money transfers	0.72	0.67	0.69	66
Mortgage	0.95	0.97	0.96	1492
Other financial service	0.75	0.27	0.40	11
Payday loan	0.64	0.65	0.64	72
Prepaid card	0.79	0.90	0.84	86
Student loan	0.92	0.88	0.90	213
accuracy			0.88	6681
macro avg	0.82	0.78	0.79	6681
weighted avg	0.88	0.88	0.88	6681

Fig. 2. CFPB-Gemini-MLP-Product-11 Classification Report

We compare the complaint classification accuracy of the Gemini-MLP-Product-11 model, with four other previously reported models, trained with the same methodology and on the same CFPB complaints dataset [35].

These include two baseline models, a bag-of-words Multinomial Naive Bayes classifier (BoW-MNB) and a bag-of-words multi-layer perceptron (BoW-MLP), and two pre-trained BERT transformer models, DistilBERT and FinBERT, fine-tuned on the CFPB complaints train dataset. Model details on training and evaluation procedures are described in [35].

Fig. 3 summarizes the performance of the 'Gemini-MLP-Product-11' model and the four previous models: For each model we show the number of parameters of the model, and its accuracy on the test and validation splits. In general, LLMs like Gemini do not always perform better than much smaller fine-tuned models, like BERT [36].

Model	Model Parameters	Test Accuracy	Development Accuracy
BoW-MNB	220,000	77.8%	79.0%
BoW-MLP	2,577,801	84.4%	86.7%
Fine-tuned DistilBERT-base	66,961,931	87.05%	86.86%
Fine-tuned ProsusAI/FinBERT	109,490,699	88.05%	87.56%
Gemini Pro-MLP no fine-tuning	7 billion (est.)	88.00%	88.47%

Fig. 3. Model parameters and Test-Development Accuracy

B. Tasks 2 and 3: Evaluation of complaint summarization and response generation

We evaluate the complaint summarization and response generation tasks on a sample of 50 randomly-selected complaints of our dataset. We use ROUGE [37], a widely used automatic evaluation metric, as well as human evaluation of machine summaries on the same complaint sample. The sample complaints range in length from 15 to 792 words.

ROUGE requires a reference summary for each complaint and uses n-gram co-occurrence statistics to compute similarity between each machine summary and the reference summary, using lexical overlap as a proxy measure. We manually wrote one reference summary for each complaint and computed ROUGE-1 (unigram), ROUGE-2 (bigram) and ROUGE-L (longest common sub-sequence, LCS) F-score statistics for each machine summary in the evaluation set. For this evaluation we use target text lengths of 25 words for the reference and machine summaries.

Additionally, given the reference summaries, one of us manually scored each machine summary, comparing it to the corresponding reference summary assigning a 3-point score, 0 (human summary better), 0.5 (similar summaries), and 1 (machine summary better). The result for the 50 summaries was 33/50 similar summaries (66%), 14/50 better human summary (28%) and 3/50 better machine summary (6%).

Evaluation of the complaint response generation task can be done similarly.

Fig. 4 shows the F-score mean and standard deviation of ROUGE-1, ROUGE-2, ROUGE-L, and of the human evaluation scores for the complaint summarization task, using our 50 complaint random sample.

Summarization Scores (F1)	ROUGE-1	ROUGE-2	ROUGE-L	Human
mean	0.2700	0.0510	0.2138	0.3939
std	0.1094	0.0680	0.0950	0.2727

Fig. 4. ROUGE and Human Evaluation Scores for Summarization

We note that the ROUGE-2 (bigram) mean score is much lower (0.051) than the ROUGE-1 or ROUGE-L mean scores, indicating low overlap of bigrams between machine and human summaries. In general, ROUGE fails to capture summary content quality [38] and the correlation between ROUGE and human evaluation scores is known to be low [39]. Those factors, apart from the need to have a reference score, are reasons for seeking alternative evaluation scores.

The evaluation of automatic text generation is thus rapidly evolving with the advent of transformer-based dense vector encodings of text (e.g., BERTScore [40]) and, in the last year, with large language models (e.g., G-Eval [41]).

C. Computation, machines, runtime and cost of LLM use

Using large language models in applications requires practical consideration of application volumes (tokens), models, computation required, runtime hardware (CPUs, GPUs, memory, bandwidth) and cost of LLM use, priced per input token and output token. Thus, we must also consider whether the tasks output text or not. In our case, we have one embedding task for classification, and two tasks that output complaint summaries and responses, with about 60 output words each per complaint (about 1/3 of input words). [3]

Google Gemini API pricing is per character, $0.000125 per input 1k characters, $0.000375 per output 1k characters [43]. For the Kaggle dataset (64M characters), about 192M input characters are processed. For the full CFPB dataset (2 billion characters), about 1.3 billion output characters are produced. This results in a cost of $40 US for processing the Kaggle dataset, and $1,255 US for the full CFPB dataset.

VI. Interpretability, Explainability and Safety

New AI technology must be developed and deployed responsibly along a number of dimensions, including model risk, bias, interpretability and explainability; as well as impacts, safety and potential for misuse.

The development and deployment of AI is a new area that is already the subject of emerging regulatory policy [19] and practice standards [44]. Thus AI systems like the one presented here, with potential for application in areas of wide impact, such as business operations and customer service, must take into account the existing regulation and upcoming standards.

Two key areas to consider for new models and applications are the use of model cards [45] and safety checking and validation of model inputs and outputs.

The Google Gemini API includes safety checks (safetyRatings) on model inputs (promptFeedback) and outputs (candidates), along a number of harm categories, each evaluated on a categorical probability scale. The safety ratings cover four categories (Harassment, Hate speech, Sexually explicit, and Dangerous), and have four probability categories (Negligible, Low, Medium and High) [46]. This functionality in the API is key to facilitate responsible product development and use of the technology, as well as the emergence of best practices.

VII. Future work

We are currently scaling up the application and runtime infrastructure for development and inference. This will enable handling larger datasets and transaction volumes, and enable choice of the large language models to be used, via commercial cloud or local customized open source models (e.g., Google Gemma, Meta LLaMA 3 [42], etc.).

We are also developing datasets and domain-specific evaluation methods, as a foundation for continuous system monitoring, evaluation and improvement. This includes client-specific business use cases that allow creating grounded evaluation methods and datasets.

While the system presented here is for textual communications only, in English, we plan to add the voice modality for input and output, as well as multi-linguality to the system (e.g., handling documents in multiple languages, and translation) — tasks to which the new generative models are well adapted.

VIII. Discussion and Conclusion

Customer communications must be serviced effectively, reliably and in a cost effective manner, at a scale that ranges into the billions of documents per year for large organizations. In this paper we presented recent advances of Generative AI, and a system for textual complaint classification, summarization and response generation, using the Google Gemini large language model. We also presented the CFPB Consumer Complaints Database for evaluation of the machine learning models for the three tasks proposed. For complaint classification, *without fine-tuning*, the accuracy is 88%, similar to one with a fine-tuned FinBERT model, which is competitive with the human level of performance. Our evaluation of automated summarization and response generation with LLMs is preliminary, but already shows competitive performance of machine vs. human reference summaries, in at least 72% of the cases. Finally, we addressed questions of AI interpretability, explainability, safety and emerging legal frameworks.

Acknowledgments

We thank our IEEE Andescon 2024 reviewers for their helpful comments. We also thank J. McCormack and G. Correa

[3] We do not consider here running the model locally; e.g., Google *Geema*, the open source version of Gemini, or Meta *LLaMA 3* [42].

for patiently listening to us discuss CFPB complaints and large language models in the past year. W. Zadrozny was partly funded by the National Science Foundation (NSF) grant number 2141124.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[6] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback, 2022," *URL https://arxiv.org/abs/2203.02155*, vol. 13, p. 1, 2022.

[7] OpenAI, "ChatGPT," https://openai.com/blog/chatgpt, 2022, accessed: March 2024.

[8] Google, "Introducing Gemini 1.5, Google's next-generation AI model," https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/, Feb. 2024, accessed: March 2024.

[9] Hugo Touvron *et al*, Meta AI, "LLaMA: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[10] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, "Large language models: A survey," *arXiv preprint arXiv:2402.06196*, 2024.

[11] E. Siegel, *The AI Playbook: Mastering the Rare Art of Machine Learning Deployment*. MIT Press, 2024.

[12] E. Mollick, *Co-Intelligence: Living and Working with AI*. Penguin-Random House, 2024.

[13] ——, "One useful thing," https://www.oneusefulthing.org/, 2024, accessed: 2024-03-26.

[14] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. Lillicrap, J.-b. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.

[15] Kaggle, "U.S. consumer finance complaints: U.S. consumer complaints on financial products and company responses," https://www.kaggle.com/datasets/kaggle/us-consumer-finance-complaints, 2016, accessed: February 2024.

[16] Consumer Financial Protection Bureau, "Consumer complaints database," 2022, accessed: February 2024. [Online]. Available: https://www.consumerfinance.gov/data-research/consumer-complaints/

[17] D. Araci. (2019, Jun.) FinBERT: Financial sentiment analysis with pre-trained language models. [Online]. Available: https://arxiv.org/abs/1908.10063

[18] S. Gopalakrishnan, "Building computational representations of medical guidelines using large language models and transfer learning," Ph.D. dissertation, The University of North Carolina at Charlotte, 2023.

[19] European Commission, "Policies, regulatory framework: AI Act," https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai, Mar. 2024, accessed: March 2024.

[20] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. LIX, p. 433–460, 1950. [Online]. Available: https://academic.oup.com/mind/article/LIX/236/433/986238

[21] C. E. Shannon, "Prediction and entropy of printed english," *Bell system technical journal*, vol. 30, no. 1, pp. 50–64, 1951.

[22] A. G. Oettinger, *Automatic language translation: Lexical and technical aspects, with particular reference to Russian*. Harvard University Press, 1960.

[23] G. Salton *et al.*, "The SMART system – retrieval results and future plans," *Information Storage and Retrieval*, pp. 1–9, 1966.

[24] J. Weizenbaum, "ELIZA — a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[25] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Computational linguistics*, vol. 16, no. 2, pp. 79–85, 1990.

[26] H. Schütze, "Part-of-speech induction from scratch," in *31st Annual Meeting of the Association for Computational Linguistics*, 1993, pp. 251–258.

[27] Y. Bengio, P. Ducharme, Réjean; Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003. [Online]. Available: https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf

[28] F. Chollet, *Deep learning with Python*. Simon and Schuster, 2021.

[29] B. Steffes, P. Rataj, L. Burger, and L. Roth, "On evaluating legal summaries with rouge," in *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, 2023, pp. 457–461.

[30] H. Zhang, X. Liu, and J. Zhang, "Summit: Iterative text summarization via ChatGPT," *arXiv preprint arXiv:2305.14835*, 2023.

[31] W. Kryscinski, B. McCann, C. Xiong, and R. Socher, "Evaluating the factual consistency of abstractive text summarization," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 9332–9346. [Online]. Available: https://aclanthology.org/2020.emnlp-main.750

[32] P. Watanangura, S. Vanichrudee, O. Minteer, T. Sringamdee, N. Thanngam, and T. Siriborvornratanakul, "A comparative survey of text summarization techniques," *SN Computer Science*, vol. 5, no. 1, p. 47, 2023.

[33] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A systematic survey of prompt engineering in large language models: Techniques and applications," *arXiv preprint arXiv:2402.07927*, 2024.

[34] B. Chen, C. Shu, E. Shareghi, N. Collier, K. Narasimhan, and S. Yao, "Fireact: Toward language agent fine-tuning," *arXiv preprint arXiv:2310.05915*, 2023.

[35] N. Correa and A. Correa, "Neural text classification for digital transformation in the financial regulatory domain," in *IEEE Andescon 2022*, Barranquilla, Colombia, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9989638/

[36] S. Gopalakrishnan, L. Garbayo, and W. Zadrozny, "Causality extraction from medical text using large language models (LLMs)," *arXiv preprint arXiv:2407.10020*, 2024.

[37] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," 2004. [Online]. Available: https://aclanthology.org/W04-1013/

[38] E. Reiter and A. Belz, "An investigation into the validity of some metrics for automatically evaluating natural language generation systems," 2009.

[39] A. Fabbri, W. Kryscinski, B. McCann, C. Xiong, R. Socher, and D. Radev, "SummEval: Re-evaluating summarization evaluation," 2021. [Online]. Available: https://aclanthology.org/2021.tacl-1.24

[40] T. Zhang, V. Kishore, F. Wu, K. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," 2020. [Online]. Available: https://arxiv.org/abs/1904.09675

[41] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-EVAL: NLG evaluation using GPT-4 with better human alignment," 2023. [Online]. Available: https://aclanthology.org/2023.emnlp-main.153/

[42] Meta AI, "Introducing Llama 3.1: Our most capable models to date," https://ai.meta.com/blog/meta-llama-3-1/, 2024, accessed: July 2024.

[43] Google, "Google Gemini pricing," https://cloud.google.com/vertex-ai/generative-ai/pricing, 2024, accessed: 2024-03-26.

[44] National Institute of Standards and Technology, "AI risk management framework," https://www.nist.gov/itl/ai-risk-management-framework, Feb. 2023, accessed: March 2024.

[45] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. Raji, and T. Gebru, "Model cards for model reporting," in *ACM FAT*'19: Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 2019, accessed: 2024-03-26.

[46] Google, "Google AI for developers, products, safety settings," https://ai.google.dev/docs/safety_setting_gemini/, 2024, accessed: March 2024.