EPiC Series in Computing

Volume 104, 2024, Pages 185–192

Proceedings of 3rd International Workshop on
Mathematical Modeling and Scientific Computing

Classification Algorithm for Creating Optimal Questionnaires in Life Sciences and for AI

Alexey P. Martyushev

Breeze Expert Limited, Auckland, New Zealand
martyushev.alexey@gmail.com

Abstract

The number of user-to-machine interactions has increased dramatically over the last decade. With the introduction of fully functional generative AI, such as ChatGPT, Copilot, and Gemini, the number of the interactions and the associated amount of information will further increase in the nearest future. This work suggests an algorithm for optimizing the number of question-reply pairs between users and a machine. The goal of the optimization was to find the optimal amount of information for API and its server. As a result of the optimization, the cost of running the services for service providers and for users can be lowered. Furthermore, in this study, the optimization led to 23% increase of customer retention rate, 15% increase of revenue per customer, 17% drop of the customer acquisition cost, and 35% increase of customer engagement with AI.

1 Introduction

Generative AI has a great perspective to become a key agent in the interaction between users and services. The AI may not be just an assistant for a search within some knowledge database, but also it can serve for the initial processing of customer requests and their classification, in apps and online [1]. Although the current versions of generative AIs are capable to produce meaningful answers to general questions [8], however, their capacity falls dramatically upon user questions become specific [2]. In particular, this effect is strong for narrowly defined services in agriculture, design, health care, and luxury goods industries.

In order to process specific user requests effectively, a generative AI model may need adjustments of its parameters to fit better to the addressed problem [3]. One may suggest that fine-tuning of model neural network (NN) parameters with extra training on a domain-specific dataset could customize the model. However, that requires to re-estimate a large number of parameters, especially for language models, such as large language models (LLMs) and small language models (SLMs) [5]. For these, the number of parameters can vary from millions to billions. This requires a significant computation power to perform the model training [4].

Meanwhile, the real complexity of AI customization may be significantly lower than the refit of million parameters. In particular, this problem is common for small businesses committed to specific domains [6]. In principle, the fine-tuning of a large model can be simplified to the introduction of an intermediate neural network of a smaller size between the users and the large

V.L. Turova, A.E. Kovtanyuk and J. Zimmer (eds.), MMSC 2024 (EPiC Series in Computing, vol. 104), pp. 185–192

model. That can fully customize a request from users at their side and then send it to a larger model, such as ChatGPT, Copilot or Gemini, for processing. This approach keeps the original AI untouched, while the customized AI (cAI) will perform decomposition, structurization, and simple transformation of user requests [11]. Furthermore, such cAIs belong to their businesses that adds extra value to their intangible assets. This is extremely valuable for the startups of early development stage.

In this study, the author demonstrates important insights into cAI implementation based on the commercial experience of Breeze Expert Limited (Auckland, New Zealand). In particular, the suggested solution was implemented for personalized mental health care and consultancy services. Furthermore, the solution discovers the psychological aspects of AI-to-user communication. For that, a created sentiment-assessment tool was monitoring the communication. The basis for psychological state assessment was browsed among a large number of factors that initially seemed promising to explain the behavior of users. However, during the proof-of-concept (PoC) experiments, the wide basis has shrunk to a narrower one while its explanatory power remained the same. Then, the basis was applied to the neural network of cAI. This refined the requests from users with a better structure, message preprocessing, and sentiment recognition.

Overall, the conceptualized approach improved the accuracy of user requests to AI. Additionally, the customized AI led to better user experience (UX), and that increased the customer retention rate. Finally, the solution optimized the operational costs of the company through the precise classification of customer needs and by minimizing the number of cases that require operator attention.

2 Materials and methods

2.1 Data collection and structure

In order to fit the parameters of cAI neural network, the data on customer-to-AI interaction was collected. The whole dataset consisted of sub-datasets collected from several trials. For the first trial – untrained model, the interaction between users and the original AI (ChatGPT) was recorded, i.e the questions of the users to AI as well as AI replies to users were recorded. The recording criteria was the daily user-to-AI interaction of 100 users with no less than 15 requests/day observed for at least one week.

In the next trial, the data of the same sample size was collected. Some of the users might not be identical to those in the first trial. The 2nd sub-dataset was randomly split at a ratio of 1:1. For this, one piece served as control, not used in training, while another one was used to refine the model parameters through extra training that resulted in 150 records in total after the second trial. For next trials, from 3th to 5th, a split ratio of 1:3 was applied to the sub-datasets of 100 size, where the smaller portion served as control.

The retrieved data includes: user activity over the measurement period, demographic data, message content sent/received to/from AI, and the emotional state of the user analyzed with the emotional basis approach – behavioral data. This measures ten distinct (non-correlating) components, see the details below.

2.2 Evaluation of AI responses

The evaluation of the quality of AI responses was performed manually by a panel group of five employees familiar with the company services. The evaluation scale ranged from 0 to 6 with the step of one, described only numerically, where 0 is the lowest quality and 6 is the highest,

or the most precise one. The resulting score was calculated as an average of four, with the dropout of a randomly selected one.

2.3 Scores of the emotional basis

The association of user message content with basis components (factors) was measured with the linguistic association between the words of the message and the words specific to the factors grouped within dictionaries. The association was numerically calculated with the similarity function of spaCy Natural Language Processing (NLP) library version 3.6 in Python. The factor magnitude equals to the sum of the spaCy association values normalized to: 1) the number of words in the message, and 2) the number of words in the related dictionary. In messages, some of the words were filtered, this includes: articles, coordinators, 'to be' verbs, subordinators, and other special grammar words of English.

2.4 Neural network of the customized AI

The neural network was designed as fully connected one with three hidden layers of 12 nodes in the first layer and of 64 nodes in each 2nd and 3rd layers. As demonstrated, this architecture guarantied effective processing of the user input messages given to cAI neural network. The scores of the emotional basis were calculated separately during message preprocessing, and then those were added into the NN input. The parameters of cAI neural network were fitted to the input data in a recurrent manner over the consecutive trials. Dropouts were applied.

2.5 User cohort and privacy

The users of this study were enrolled to the test trials after the plan and the goals of this study were disclosed to them. The users agreed and signed the participation (volunteer) agreement. Their personal data was used only for the purposes listed in the plan of this study.

2.6 Metrics

The standard metrics of user experience as well as the customer retention rate were calculated in order to evaluate the effect of customized AI on the quality of company services [6]. The exact numbers for each category cannot be disclosed here because of the privacy requirements of the company, however, the aggregate numbers are provided.

2.7 Coding and computational experiments

The solution was programmed in Python version 3.8 and PyTorch version 2.3.0. The Python code was used for the initial data processing, structurization, and visualization with standard Python libraries that include: Matplotlib version 3.8.1, Pandas version 2.1.0, Scikit-learn version 1.4.2, and WordCloud version 1.8.1. The PyTorch code was used to program and train the neural network. The simulations were performed on a local machine with MS Windows 11, CUDA version 12.1, Intel Core i9-13900K, 32 GB RAM, NVIDIA GeForce RTX 4090, and 2TB SSD. Further information is available upon request.

3 Results

The concept development, tests, and validation were done in several steps. Most of the work was done on the integration of the data into the cAI model (Fig. 1).

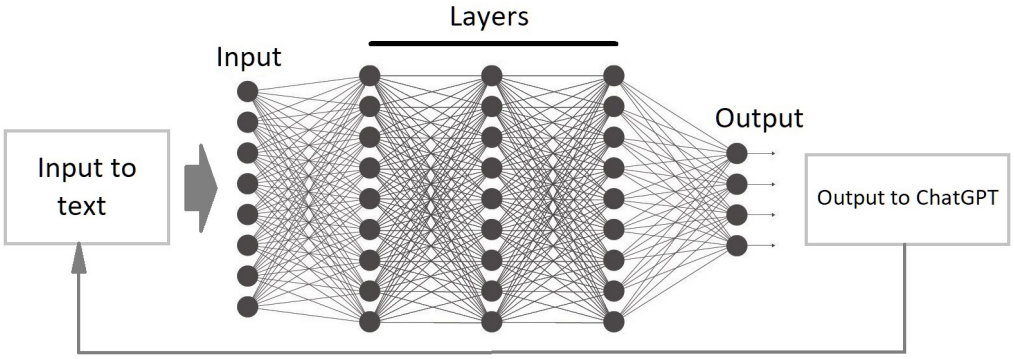

Figure 1: An abstraction of the applied architecture.

First, it required to clean and sort the raw data into categories with a certain criteria that are the most effective for the considered services. For example, a user request to AI can be split into several bins with specific keywords or some particular structures of message sentences. Then, the bin contents can be processed through the first hidden layer of the cAI network.

Second, the model parameters were fitted to the data in a recurrent manner that allowed us to use the data effectively and utilize the memory effect of the network. The data of consecutive trials served for that, see in Materials and Methods.

Third, the model parameters were tested on various types of variation. For this, the confidence intervals were estimated with a general formula and bootstrapping, while dropouts were applied. Some nodes with their parameters demonstrated an extremely high variance, and therefore these nodes were pruned. Later, the analysis has shown that the model did not lose in accuracy while the tolerance has increased.

Finally, at the validation step, the fully-assembled model was tested with two extra trials, 6th and 7th ones. The data for these trials was collected consequently with a time gap of three months after the data of the 5th trial was collected. This means that the adaptation of users to the improved company services may provide a bias. The actual scale of this effect was not tested directly, however, the performance of the model on those datasets did not deviate significantly from the values observed for the test datasets.

3.1 Data cleaning

After the data was collected, as described in Materials and Methods, it was checked on the presence of invalid records and fake communication. The fake communication here means the requests to AI that are either unrelated to the company services or meaningless as a whole.

As a result of cleaning, 11% of data grouped by user ID were cleared. The remaining data did not demonstrate anything suspicious, and the user-to-AI communication was meaningful. The demographic data was also used in the model training. The inputs to NN were processed even when some categories of the demographic data were omitted.

The responses of AI were also collected and used for evaluating AI performance. Data check did not reveal any low-quality responses, however, some of the responses were imprecise and unrelated to the topic. Thus, only 11% of customer messages were neglected, with the corresponding AI responses. Then, the remaining data was used for fitting the NN parameters and to calculate their confidence intervals.

3.2 Model setups and tests

Using the data, the model parameters were fitted in PyTorch with Adam method. The speed of convergence was observed intermediate. So, the method did not require any extra actions on the step parameter. Dropouts did not change dramatically the dynamics of convergence, except a few cases, when the error curve leveled off after 50-70 iterations for some epochs. The number of epochs was set to 20. The confidence intervals of the parameters were calculated manually with the common formula in PyTorch.

After the model parameters were fitted, the model was run with the estimated parameters on a test dataset, which comprises 20% of the original data. The main purpose of the test was to assess how the model performs on the datasets that it was not trained on. The test dataset demonstrated improved AI responses in 65% cases when the customized model was compared with the original one. This performance is similar to that observed for the training datasets. In 90% cases, the cAI responses were not worse than the original ones that is again similar to that observed for training datasets. The quality of the responses was assessed by the same expert panel, as described in Materials and Methods. To reproduce a blind experiment, the responses produced with the cAI model were mixed with the responses from the original model, at a ratio of 1:1, and then those were given to the panel for evaluation. The time gap between the assessments was four months.

At the validation step, the metrics were tested again on additional 20% of data. The observed deviation of metric values from the test datasets was less than 5% that is good. We concluded that the quality of the model fits to client requirements, and the model was not further improved.

3.3 Analysis of cAI response quality

The quality of AI responses was comprehensively analyzed in order to find possible groups of responses for which the cAI performed significantly differently than was observed for the original AI. For that, the responses created by AIs were compared on the test and validation datasets. The improvement in AI response quality was measured quantitatively as the average increase of quality scores evaluated by the expert panel.

Among the responses, three distinct groups were observed. The groups differ in the topic of the requests that include: 1) general questions on health issues, 2) the methods of treatment, and 3) the requests related to the emotional support of clients. Within each of the groups, some subcategories can also be identified in order to split the groups into smaller subgroups, but it was out of the scope of this study.

For the three groups, the average improvement of the quality of AI responses was calculated. The results did not demonstrate any significant increase of the response quality for the group of general health questions. However, the cAI quality was significantly higher for 'methods of treatment' and 'emotional support' groups. The higher quality was likely associated with the specificity of the subjects and related extra training performed for cAI in order to get it familiar with the subjects. So, the requests to AI related to a broader topic, here it was the general health questions, are likely to be answered satisfactory with the original AI, without

any customization applied [3]. This makes sense, when the abundant information on the topic is available on the Internet. However, once the question specificity increases, especially when the questions are directly related to the company services, the quality of the responses drops dramatically for the non-customized AI.

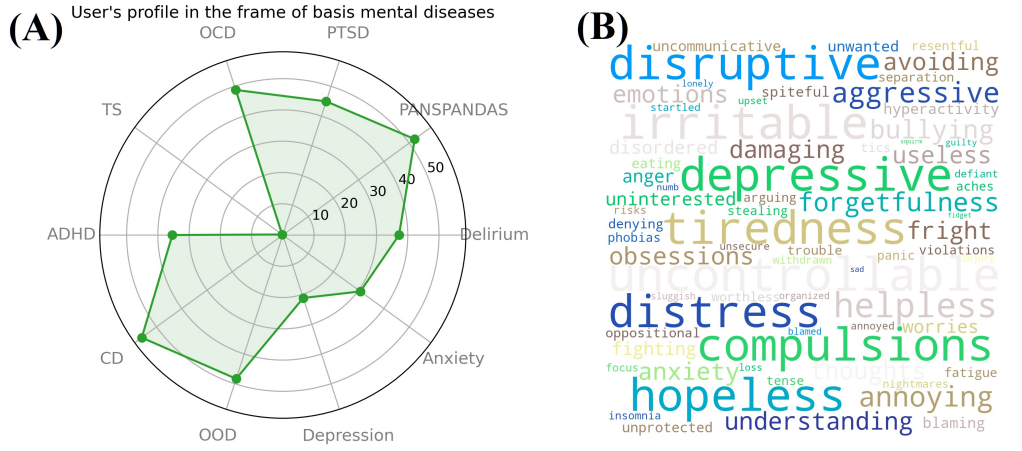

Figure 2: (A) Emotional basis and (B) the associated words.

Additional attention requires 'emotional support' group of questions since it received significantly better scores with the customization. The reason for this difference likely lies within the fundamental aspects of machine learning methods that are different between the two compared AI systems. Potentially, the assessment of sentiments is not a proper use of the standard ChatGPT engine, while with the customization (Fig. 2), higher scores can be achieved. Meanwhile, the increase of the quality scores might be an effect of the bias contributing at several steps of cAI implementation. In conclusion, sentiment-related tasks, such as we considered, likely will require supervised extra training in order to reach a higher quality of responses [7].

3.4 The effect of cAI on client services

A commercially oriented project has to demonstrate a positive effect on company profit in order to justify its implementation. This criterion applies to our work as well, after the suggested R&D solution was implemented for the client company that focuses on end users.

In order to evaluate the economic effect of developed cAI, first, the retention rate of customers was compared before and after the launch of cAI in services. The 3-month time point after the cAI launch was set for the measurement. The retention rate was calculated on the traffic data of users who used the service for four weeks since the beginning of measurements. During that period, the retention rate was measured for new customers who began using the company services during the measurement period. For the 'before the launch of cAI' group, the measurements were collected during the period of the same length. In order to adjust the retention rate to the number of customers acquired during the measurement period, a scaling factor was applied.

The results demonstrated a 23% increase of the customer retention rate within the four-week measurement period. The company management considered this result as satisfactory. Although that number may be limited to represent completely the effect of AI customization.

This is because the potential bias of other changes could be present during the measurement period. Other measurable changes were not significant, for example, ones that are related to marketing initiatives and promotions, however, for a developing business, even small changes may be significant. In addition to the retention rate, the economic effect of cAI launch was measured with other metrics. For example, the amount of sales per customer has increased by 15%. Furthermore, the cost per customer acquisition has dropped by 17%. At the same time, the customer engagement on social media almost did not change after cAI launch, while the user engagement (time) with AI has increased by 35%.

Overall, the results demonstrate a positive effect of cAI launch on the small business, for which a constant customer acquisition rate has already been established. Meanwhile, the market expansion was on a relatively small scale when compared with main competitors.

4 Discussion

Although the recent advantages in the development of generative AI demonstrate fascinating results in producing meaningful contents, however, the limitations are obvious. These limitations come mostly from the imperfectness of neural network architecture. That includes a large number of model parameters for fitting during model training, the determinism of signal propagation within the networks, and poor performance on small training datasets [4]. For AI, in order to reply properly to the diverse requests of a relatively small number of users, it is better to customize the model network with an extra network - cAI that has a significantly less number of hidden layers, and therefore a less number of parameters for fitting. This study considers such cAIs and their application to client-focused businesses.

The results of this study demonstrate a significant increase of both customer engagement to AI services and UX quality, after the cAI solution was launched. Those two metrics were initially targeted for improvement. This demonstrates how R&D can precisely and effectively meet businesses needs in a project form within certain time frames. Similar approaches researched online did not demonstrate such effectiveness. Therefore, there is a significant market gap between the large-scale AI solutions and customized ones.

Technically, the implementation process was quite standard for neural network engineering with parameter estimation. However, the preliminary analysis of data and sorting this into categories were innovative and differ from similar approaches. In particular, the sentiment recognition with the suggested emotional basis is a unique solution that can be protected by intellectual property rights. Overall, this work demonstrates a significant impact of small solutions targeting business needs as a complementary product to the global large-scale AI solutions [10]. Furthermore, this project demonstrates that the creation of intellectual property can be localized at small businesses, for meeting their specific needs. That can help to promote the business to a larger share in targeted markets.

Among the possible caveats of this study, there is a possible bias of sentiment basis selection (Fig. 2) for identifying the psychological states of customers. Due to the basis selection was data-driven, the input data has a direct effect on the results. Therefore, the data sample parameters such as its size and quality could affect the results. Additionally, the parameters of cAI were evaluated through fitting to the data that ultimately adds bias. Finally, some data categories include the scores evaluated by the expert panel that might be subjective due to the nature of the experiment [9].

In conclusion, the offered customized AI system (Fig. 1) improves the requests of users to AI, while the complexity of the system remains relatively low that allows us to modify and tune the solution easily upon request. That model setup produces meaningful results and it satisfies the

current company needs, such as the retention of customers and a potential expansion within the market. Finally, an extra training of the cAI can further improve its performance and accuracy.

References

[1] M. A. Abu Daqar and A. K. A. Smoudy. The role of artificial intelligence on enhancing customer experience. *International Review of Management and Marketing*, 9(4):22–31, 2019.

[2] N. Cong-Lem, A. Soyoof, and D. Tsering. A systematic review of the limitations and associated opportunities of ChatGPT. *International Journal of Human–Computer Interaction*, pages 1–16, 2024. https://doi.org/10.1080/10447318.2024.2344142.

[3] L. De Angelis, F. Baglivo, G. Arzilli, G. P. Privitera, P. Ferragina, A. E. Tozzi, and C. Rizzo. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health*, 11:1166120, 2023.

[4] Epoch AI. Key trends and figures in machine learning. https://epochai.org/trends, Accessed: November 2024, 2023.

[5] J. Gerstmayr, P. Manzl, and M. Pieber. Multibody models generated from natural language. *Multibody System Dynamics*, 62:249–271, 2024.

[6] Y. Gupta and F. M. Khan. Role of artificial intelligence in customer engagement: a systematic review and future research directions. *Journal of Modelling in Management*, 19(5):1535–1565, 2024.

[7] J. K. Kim, M. Chua, M. Rickard, and A. Lorenzo. ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. *Journal of Pediatric Urology*, 19(5):598–604, 2023.

[8] OpenAI. ChatGPT: large language model. https://chat.openai.com, Accessed: November 2024, 2024.

[9] H. Park, A. Megahed, P. Yin, Y. Ong, P. Mahajan, and P. Guo. Incorporating experts' judgment into machine learning models. *Expert Systems with Applications*, 228:120118, 2023.

[10] I. Sarker. AI-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems. *SN Computer Science*, 3(2):1–20, 2022.

[11] S. Song and S.-W. Lee. A goal-driven approach for adaptive service composition using planning. *Mathematical and Computer Modelling*, 58:261–273, 2013.