# Augmenting Persuasive Argument Datasets using Large Language Models*

Renan Lirio de Souza[1,2,*,†], Mauro Dragoni[2,†]

[1]Free University of Bolzano (UNIBZ), piazza Università 1, 39100 Bozen-Bolzano, Italy
[2]Fondazione Bruno Kessler (FBK), Via Sommarive 18, 38123 Trento, Italy

## Abstract

Data augmentation comprises a set of techniques that are used to improve the performance of machine learning models. In the Natural Language Processing (NLP) field, the discrete nature of language represents a challenge to data augmentation, as text can lose coherence or syntactic accuracy during the augmentation process. This challenge is particularly pronounced in persuasive dialog systems, where high-quality data is crucial and scarce due to privacy regulations. In this study, we investigate the application of Large Language Models (LLMs) to enhance persuasive arguments from an Automatic Persuasive System (APS). Using a limited COVID-19 dialogue dataset of user-machine persuasive interactions, our goal is to evaluate the efficacy of different augmentation techniques, combined with LLMs, in generating syntactically coherent and accurate dialogues, with a specific emphasis on dialogue quality and persuasiveness. The results show that integrating LLMs with augmentation generates realistic and diverse examples, aiding the overall quality and effectiveness of the persuasive dialogues produced by the system.

## Keywords

Data Augmentation, Natural Language Processing, Large Language Models, Persuasive Dialog Systems, Arguments

In the realm of digital communication, Automated Persuasion Systems (APS) serve as digital interlocutors, assuming the role of a persuader to influence users (persuadees) through compelling arguments toward achieving predetermined objectives [1]. This conversation system requires a large number of interactions to operate effectively, which can be defined as labeled high-quality conversational (dialogue) data, a resource often scarce, especially in sensitive domains such as healthcare [2]. However, collecting these interactions is a challenging and resource-intensive process that involves the creation and implementation of surveys, identifying participants, annotating data, and managing privacy concerns [3].

One common approach to address data scarcity issues is applying text-data augmentation (DA) techniques. DA involves generating new text data from existing samples, without the need for additional real-world data collection [4]. This process can enhance the volume and diversity of training datasets and improve the performance of the model in NLP tasks such as text classification, machine translation, and dialogue generation by generating additional samples with label-preserving transformations [5]. However, traditional DA strategies often rely on simple textual transformations, such as random word replacement, addition, removal, or swapping [6]. Although these methods can enhance model robustness against minor variations, they may not be sufficient to generate good persuasive arguments.

The emergence of Large Language Models (LLMs) such as GPT [7] has introduced a new era of possibilities for textual DA. Trained on vast and diverse datasets, these models excel at generating coherent and contextually appropriate text across various domains and can be tailored to act as possible data augmentators. The main goal of my PhD is to develop and evaluate a method that implements LLM's capabilities to generate novel and contextually accurate persuasive dialogues and arguments for an APS system. We applied this method as a case study using a COVID-19 dataset of user-machine persuasive interactions, composed of user arguments and concerns. In this paper, we used a small

*Doctoral Consortium at the 23rd International Conference of the Italian Association for Artificial Intelligence Bolzano, Italy, November 25-28, 2024.*
*Corresponding author.
†These authors contributed equally.
rliriodesouza@fbk.eu (R. L. d. Souza); dragoni@fbk.eu (M. Dragoni)

sample size of arguments from the COVID-19 dataset to test and evaluate the method by applying three different augmentation approaches: paraphrasing, backtranslation, and masking.

## 1. Dataset

The dataset used to evaluate our approach is based on the work of [3]. They developed and provided a knowledge base consisting of a set of aggregated arguments and counterarguments related to common concerns about COVID-19 vaccination. From the available datasets, we choose the *concern_classifier_new_na* dataset, which is composed of tuples of arguments ($A$) and concerns ($C$). We defined it as $D_b$, as presented in Table 1, where:

$$D_b = \{(id_i, A_i, \mathrm{C}_i)\}_{i=1}^{N_b}$$

with each tuple ($N_b$) consisting of a unique identifier ($id$), an argument ($A$) and a concern ($C$), where:

- $id_i$ represents the unique identifier for the $i^{th}$ argument.
- $A_i$ contains a user interaction of the $i^{th}$ argument, which includes a user's dialogue against getting a COVID-19 vaccine.
- $C_i$ is the category of concern that the $i^{th}$ argument addresses, such as side effects, vaccine development speed, efficacy, safety, etc.
- $N_b$ is the total number of arguments in the base dataset.

| id | Concern | Argument |
|---|---|---|
| 1 | healthy | I have strong immunity |
| 2 | healthy | I am not worried about my health anyway |
| 3 | healthy | I am healthy and not worried to catch COVID |
| 4 | healthy | I know that, but I'm in good health for my age |
| 5 | healthy | I'd rather let my immune system develop naturally |
| 6 | healthy | I prefer my chances against Covid 19, I am young and healthy |
| 7 | healthy | True but I am young and healthy and I don't think I will need the vaccine |
| $id_i$ | $C_i$ | $A_i$ |

**Table 1**
Small sample size of the dataset $D_b$, highlighting some examples of the healthy concern. These examples were used to evaluate the proposed augmentation idea.

We randomly selected these 7 arguments of the *healthy* concern to be used as original data for augmentation purposes and evaluated our proposed augmentation method. We also selected only 7 arguments because of the limited number of pages allowed for the paper, as more selected arguments could lead to more augmented output examples. As the main idea is to increase the number of existing examples in $D_b$ with synthetic arguments, we choose the *healthy* concern as it is a class with a total amount of values in the middle compared with other concerns, as presented in 2.

The complete dataset comprises, in total, 820 user arguments ($A$), divided into 15 distinct concerns ($C$). Table 2 also provides a breakdown of the frequency of each concern, measured in terms of the number of arguments associated with it. The most common concerns are "long-term effects" (19%), "safety" (17%), and "side effects" (15%). The dataset presents a clear imbalance in the distribution of arguments, a common and well-documented issue in the literature. Data augmentation (DA) techniques offer an effective solution to address this imbalance, especially in improving representation among the underrepresented classes.

## 2. Method

Traditional DA techniques typically use a simplistic strategy that modifies a given sentence to improve classification algorithms' generalization capabilities [8]. While this approach is effective to a certain

| # | Concerns | Arguments | Average |
|---|---|---|---|
| 1 | long_term_effects | 154 | 19% |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 7 | healthy | 47 | 6% |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 15 | already_had | 11 | 1% |
| **Total** | **15** | **820** | **100%** |

**Table 2**
Breakdown of the COVID-19 vaccines user arguments dataset $D_b$.

extent, it may lack the ability to generate novel coherent data for humans [9]. However, this paradigm is changing with the advent of Large Language Models (LLMs). Researchers are currently using LLMs to develop novel augmentation strategies tailored to the challenges of the NLP domain [7, 10]. The study in [7] uses LLMs like ChatGPT for text augmentation by rephrasing sentences into semantically distinct variants, yielding improvements even in few-shot learning scenarios. In [10], GPT-generated samples enhanced the classification of underrepresented vaccine hesitancy in Dutch social media, and with back-translation, improved model accuracy and F1 scores. These findings show that GPT models can create realistic, diverse examples, enhancing training in imbalanced datasets.

Although LLM models have shown good performance in data augmentation, their performance varies with different datasets and tasks [11]. These advancements have expanded the possibilities for data augmentation applications, enabling more advanced and efficient approaches. However, more research is needed to establish standardized practices, especially for complex tasks such as argument augmentation for persuasive dialog systems.

In response to the lack of DA methods for persuasive argumentation, we investigated the use of LLMs to generate synthetic arguments. By exploiting the advanced capabilities of these models, we aim to develop diverse and contextually relevant arguments to enhance $D_b$. Several models were evaluated, including GPT, Gemini, BERT, and DistilBERT, but we are going to use only GPT. As a result, we formulated two specific prompts to interact with these models, as shown below.

## 2.1. Prompts Layout

We designed three different prompts to interact with the LLM, by organizing them following a structure (Figures 1a and 1b) with three main sections: Instruction, Examples, and Output.

*###Instruction:*
Create 2 sentences for each example using paraphrasing augmentation.

*###Examples (Healthy concern):*
1. I have strong immunity
2. Not worried about my health anyway
$id_i$. $A_i$ (Table 1)

(a) $P$ and $B$ augmentation prompt.

*###Instruction:*
Create 5 new sentences different from the examples, following the ###Output logic.
*###Examples (Healthy concern):*
1. I have strong immunity
2. Not worried about my health anyway
$id_i$. $A_i$ (Table 1)
*###Output:*
<mask> <mask> healthy <mask> <mask> fate <mask>

(b) Masking augmentation prompt.

**Figure 1:** Prompts templates with their respective instruction sets for sentence augmentation strategies: paraphrasing and backtranslation (Prompt 1) and masking (Prompt 2).

- **Instruction.** Main description or information for the LLM executing a task. The prompt 1 (Figure 1a) has a unique parameter that is highlighted as underline, which was used with**paraphrasing** (*P*) and **back-translation** (*B*).

- **Examples.** Examples of **arguments** from $D_b$ to provide the model with a comprehensive data context. During our tests, we used seven arguments from healthy concerns (Table 1).
- **Output.** This section provides the template or structure (in the form of a sentence with blanks) that the model should follow when creating the augmented sentences. The placeholders, represented by "<mask>", indicate where the model needs to fill in words that will create a coherent sentence. In the example provided in Figure 1b, the model needs to fill the <mask> values around **"healthy"** and **"fate"** with words that match the tone and meaning of the examples. We used these two words with the <mask> values to have more control over sentence length, generation process, and contextual integrity. The word **"healthy"** was chosen because it addresses the concern *C* itself. Conversely, the word **"fate"** was randomly selected from a bag of words with partial relevance to the topic, while also introducing subtle variation into the final sentence and forcing the model to follow the implied concept.

## 3. Experiments

We execute all prompts using Python code using the API access for GPT model. However, all prompts can be easily executed via the standard web browsers for GPT or Gemini. In this section, we will present the results obtained using the GPT model. The sentences resulting from the first two prompts (Figure 1a, with paraphrasing and backtranslation), are shown in Tables 3 and 4.

| # | Paraphrasing (*P*) |
|---|---|
| 1 | My immune system is quite robust |
| 2 | I don't have any concerns about my health |
| 3 | I'm in good health and not concerned about getting COVID |
| 4 | I understand, but I'm in good shape for my age |
| 5 | I'd rather let my body build its immunity naturally |
| 6 | I feel confident about facing COVID-19 since I'm young and fit |
| 7 | That's true, but I'm young and healthy, so I don't think I'll need the vaccine |

**Table 3**
Examples of paraphrased sentences generated by using Prompt from Figure 1a.

| # | Backtranslation (*B*) |
|---|---|
| 1 | I possess strong immunity |
| 2 | I'm not concerned about my health anyway |
| 3 | I'm healthy and not worried about getting COVID |
| 4 | I know that, but I am in good health for my age |
| 5 | I prefer to let my immune system develop naturally |
| 6 | I prefer my chances against COVID-19, I'm young and healthy |
| 7 | It's true, but I'm young and healthy, and I don't think I'll need the vaccine |

**Table 4**
Examples of backtranslation sentences generated by using Prompt from Figure 1a.

A qualitative comparison of the tables with the original sentences reveals distinct variations for each technique. The paraphrased sentences show minor changes, adopting a conversational tone, while retaining the original meaning. In contrast, backtranslation results in sentences with negligible changes. This can be attributed to the choice of language, suggesting that a combination of different languages may yield better results. The results of the masking augmentation, as presented in table 5, show significant variation in both structure and meaning, incorporating an element of randomness while maintaining the integrity of individual words. The generated sentences are also similar in length and include both predefined words. The terms "healthy" and "fate" have been added to the table (last two inputs) as "fake" sentences to evaluate the impact that each word has when compared with the original sentences.

| # | Masking (*M*) |
|---|---|
| 1 | I trust being healthy and believe my fate will protect me |
| 2 | I am staying healthy and leaving my fate to nature's course |
| 3 | Remaining healthy is important, but I accept my fate either way |
| 4 | I focus on staying healthy and let fate decide the rest |
| 5 | Being healthy gives me confidence to leave fate as it is |
| 6 | healthy |
| 7 | fate |

**Table 5**
Examples of augmented sentences generated by using masking (*M*) Prompt from Figure 1b.

We used the embedding similarity metric to assess the degree of similarity between all sentences, using the *"paraphrase-MiniLM-L6-v2"* sentence-transformers model [12]. This model maps sentences to a 384-dimensional vector space, which enables semantic comparison between sentences in all tables. It is suitable for clustering, semantic search, and other tasks.

Initially, we obtained the embeddings of the original arguments (Table 2) as well as the embeddings of the masked results (Table 5), and performed a comprehensive comparison of all sentence embeddings. The resulting similarity matrix, shown in Table 6a, provides a complete comparison between the original arguments and the corresponding masked versions. Since the masked results are not generated on a one-to-one basis (i.e., original sentence 1 to masked sentence 1), the comprehensive matrix provides a more detailed understanding of the similarity results. Paraphrasing and back-translation (Table 6b) only need diagonal entries of the matrix; as they achieve modifications at the sentence level, hence, it is not necessary to perform evaluations with the others.

| $M$ \ $D_b$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 | 0.34 | 0.25 | 0.30 | 0.13 | 0.22 | 0.36 | 0.08 |
| 2 | 0.41 | 0.44 | 0.45 | 0.43 | 0.40 | 0.50 | -0.02 |
| 3 | 0.39 | 0.46 | 0.45 | 0.47 | 0.30 | 0.50 | 0.07 |
| 4 | 0.42 | 0.46 | 0.43 | 0.40 | 0.41 | 0.56 | 0.04 |
| 5 | 0.37 | 0.34 | 0.38 | 0.34 | 0.35 | 0.30 | 0.20 |
| 6 | 0.32 | 0.36 | 0.37 | 0.39 | 0.25 | 0.31 | 0.11 |
| 7 | 0.40 | 0.29 | 0.34 | 0.27 | 0.31 | 0.37 | 0.10 |

(a) Full matrix of $D_b$ and Mask $M$

| $D_b$# | $M$ | $P$ | $B$ |
|---|---|---|---|
| 1 | 0.34 | 0.78 | 0.98 |
| 2 | 0.44 | 0.79 | 0.92 |
| 3 | 0.45 | 0.84 | 0.92 |
| 4 | 0.40 | 0.77 | 0.98 |
| 5 | 0.35 | 0.86 | 0.91 |
| 6 | 0.31 | 0.83 | 0.98 |
| 7 | 0.10 | 0.94 | 0.93 |

(b) Main diagonal

**Table 6**
Main diagonal values of embedding similarity between original dataset $D_b$ with the mask, paraphrasing 1 (P1) and paraphrasing 2 (P2) augmented data

The higher the value, the more similar the two sentences are in meaning, with values closer to 1 indicating strong similarity and values closer to 0 (or negative) indicating little or no similarity. In Table 6a, the majority of values are relatively low, typically ranging between 0.25 and 0.50. The masked sentences replace specific components of the original sentences with abstract terms, resulting in more diverse and creative outputs that significantly differ from the original sentences. However, despite the low similarity scores of these sentences, it is noteworthy how similar they are to the results obtained for the sentence using only the term "healthy" (6 in *M* table). This suggests that the masked sentences remain relevant to the main concern of the example (healthy), having different arguments within the same category. Conversely, the sentence "fate" (7 in *M* table) had the lowest similarity scores, which is not surprising since "fate" is not consistently associated with a health-related context. This also implies the potential for even lower similarity values in complete sentences, since they all contain this word.

The last table T (Table 6b) presents the results of the diagonal embedding similarity compared to the original $D_b$. Both paraphrasing and backtranslation scores show much higher similarity than masking *M*. The *P* method yields results ranging from 0.77 to 0.94, making it ideal for varied, yet faithful sentences.

The *B* method produces an almost identical structure with the highest similarity scores, providing a safer technique for preserving sentences within the same class. Finally, masking introduces more randomness and diversity, while providing control by allowing users to define specific keywords or masks to generate more varied arguments.

## 4. Conclusion and Discussion.

In this study, we explored the integration of Large Language Models (LLMs) with data augmentation techniques to improve the quality and effectiveness of persuasive dialogues generated by an Automated Persuasion System (APS). By applying methods such as paraphrasing, backtranslation, and masking to a limited COVID-19 dialogue dataset, we were able to produce syntactically coherent and contextually accurate examples. These techniques not only could address the data scarcity issue but also improve the overall effectiveness of the APS. Future work will focus on refining these approaches and exploring additional augmentation strategies like user behavior, profile and clustering by types of users.

## References


[1] A. Hunter, Computational persuasion with applications in behaviour change., in: COMMA, 2016, pp. 5–18.

[2] M. Bayer, M.-A. Kaufhold, C. Reuter, A survey on data augmentation for text classification, ACM Computing Surveys 55 (2022) 1–39.

[3] L. Chalaguine, A. Hunter, Addressing popular concerns regarding covid-19 vaccination with natural language argumentation dialogues, in: Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 16th European Conference, ECSQARU 2021, Prague, Czech Republic, September 21–24, 2021, Proceedings 16, Springer, 2021, pp. 59–73.

[4] S. Li, X. Ao, F. Pan, Q. He, Learning policy scheduling for text augmentation, Neural Networks 145 (2022) 121–127.

[5] J. Chen, D. Tam, C. Raffel, M. Bansal, D. Yang, An Empirical Survey of Data Augmentation for Limited Data Learning in NLP, Transactions of the Association for Computational Linguistics 11 (2023) 191–211. URL: https://doi.org/10.1162/tacl_a_00542.

[6] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Tepper, N. Zwerdling, Do not have enough data? deep learning to the rescue!, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 7383–7390.

[7] H. Dai, Z. Liu, W. Liao, X. Huang, Y. Cao, Z. Wu, L. Zhao, S. Xu, W. Liu, N. Liu, et al., Auggpt: Leveraging chatgpt for text data augmentation, arXiv preprint arXiv:2302.13007 (2023).

[8] Z. Hu, R. K.-W. Lee, N. F. Chen, Are current task-oriented dialogue systems able to satisfy impolite users?, arXiv preprint arXiv:2210.12942 (2022).

[9] H. Queiroz Abonizio, S. Barbon Junior, Pre-trained data augmentation for text classification, in: Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I, Springer, 2020, pp. 551–565.

[10] J. Van Nooten, W. Daelemans, Improving dutch vaccine hesitancy monitoring via multi-label data augmentation with gpt-3.5, in: Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, July 2023; Toronto, Canada, volume 1, 2023, pp. 251–270.

[11] F. Piedboeuf, P. Langlais, Is chatgpt the ultimate data augmentation algorithm?, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 15606–15615.

[12] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: http://arxiv.org/abs/1908.10084.