Human values behind arguments: Study on accuracy on large language models and strategies approaches

Sebastian Pellizzari and **Peter Burger** and **Elbaraa Elsaadany** and
Hamed Homaeirad and **Jannis David Voigt**
University of Innsbruck, Austria

Abstract

This paper delves with the capability of large language models to understand the actual meaning of a sentence provided by a human user. Imagine you write the following to an arbitrary large language model: "I have butterflies in my stomach, what should I do?". It will probably act empathically, as it has retrieved the encapsulated human values behind that sentence correctly and will provide you with a response a normal human being would provide you with.
As the area of large language models (LLMs) is currently trending, this paper is trying to contribute to this area of science by testing the accuracy of popular large language models such as ChatGPT 3.5 as well as Llama2-7b which were recently introduced to interact with its users in some way and to answer general questions rather than to classify sets of data records. For this reason, we have done an exploratory data analysis (EDA) on the given dataset [1] which has been used in (Kiesel et al.).
To do so, we have first passed the prompts which have been specifically tailored for the respective individual general LLMs, as they were not fine-tuned on this particular task and rather work on a conversational way and not solely by passing the respective triplets. Further the responds were automatically analysed using a pipeline, which can be accessed in the GitHub repository provided in the footnotes.[2] In order to compare the results of all large language models, we have used the provided dataset of (Kiesel et al.).
After carrying out various experiments with the large language models Llama2-7b as well as ChatGPT3.5 as well as running a full analysis of the provided dataset the investigation revealed that ChatGPT3.5 (average F1-score was about 0.28 for granularity level 3) performed much better than Llama2-7b (average F1-score was about 0.09). However, the F1-score for both ChatGPT3.5 and Llama2-7b were well below the fine tuned-models from (Kiesel et al., 2023) where the top models had a F1-score of around 0.55. In both cases it has been discovered that role-based prompting did hardly make any difference in all experiments. However, by using a coarser granularity (by grouping the labels to groups of similar labels) the F1-score improved drastically for ChatGPT3.5: 0.27 for finest granularity with 20 different labels, 0.39 for the next coarser granularity with twelve different labels, and 0.8 for the coarsest granularity with 4 label classes.
While executing the exploratory data analysis, it has been discovered that there are plenty of inconsistencies in all parts of the dataset (training-, validation- and test-part). For instance, there has been an entry in the dataset which had not been assigned any label. Furthermore, there have been 156 (of 8865) duplicate entries found in the dataset, with some of them having different labelling.

Nowadays, large language models (LLMs) are very clever in understanding what the user wants to express. In a further sense, LLMs are not taking the literal meaning of a sentence but they are first scanning the given sentence for certain keywords to adequately compose an answer which would make sense in the given situation. For instance, if one would tell a large language model such as ChatGPT3.5: “I have butterflies in my stomach”. the language model would act emphatically and ask if that person wanted talk about it to make them feel better or respond in a playful way. However, it would not take the literal meaning of the sentence that someone has insects inside one’s body.
This task was introduced by Kiesel at al. at SemEval-2023, which represented the fourth task. In this task roughly 9000 arguments were provided for training and testing purposes which contained short phrases and the large language models had to assign certain labels to those sentences. There were 20 value categories, which were further grouped into subgroups. This task as well as the labels will be further introduced in the subsequent chapter.

[1] Dataset: `webis/Touche23-ValueEval`
[2] GitHub Code Repository

While this problem was solved by several research groups as well as the accuracy briefly summarized with the F1 score in Kiesel et al. (Kiesel et al., 2023), there have not been any papers written which compare the LLMs that were fine tuned for this task with the ones that have a more general purpose, such as ChatGPT 3.5 Llama2.

We are trying to close this research gap by answering the following research questions:

- Is there any inconsistencies in the dataset? Is there any connection between the given labels, i.e., does the occurrence of a certain label imply another label as well being present?
- Will the LLMs provided in Kiesel et al. (Kiesel et al., 2023) which were fine-tuned for this exact task outperform general LLMs which were trained on an immensely larger training set as well as with a larger training depth?
- Will the result be any different if role-based prompts are used? (e.g., prompts asked from the perspective of a scientist or a psychologist)

The first research question will be tackled by doing an exploratory data analysis (EDA), to get familiar with the given dataset as well as with the purpose of finding any potential inconsistencies which might cause issues with the training of any LLM based on this dataset.

The paper first introduces the essential terminology to be able to understand the subsequent chapters and the given task. Then the paper continues with a thorough exploratory data analysis (EDA) which generates some general statistics, e.g., the average label assigned, average length of the given arguments and the most common words in the dataset. Further the features are analysed, e.g., which words are connected to a certain label, how many samples are assigned to a certain label, and if there is any correlation between the labels. In the last chapter the results of the experiments are discussed, where the previously mentioned LLMs are labelling the arguments from the given dataset to be able to compare the F1 score of the tested large language models with the F1 score of the LLMs which were competing in Touchés task in (Kiesel et al., 2023).

1 Background

Kiesel et al. introduced Touché's task in (Kiesel et al., 2023) which had the goal to identify human values behind arguments, which was hosted as Task 4 at SemEval-2023. For this task a dataset of 8865 arguments was given as well as a supplementary dataset of 459 arguments, which was used by all 29 participants of this study to evaluate their submissions comparably.

In the same paper (Kiesel et al., 2023) this task was introduced. The trained model should determine whether a given textual argument resorts to a given human value or not. To make the outcome of the developed models possible, 20 value categories were already given in (Kiesel et al.). In this source each of those values were grouped into three distinct granularity levels. The coarsest level (which we will from now on describe with level 1 granularity) is divided in four value categories, i.e., openness to change, self-enhancement, conservation and self-transcendence. In the overview, below we have denoted this granularity without any bullet point. The values with a better granularity are the respective bullet points which are directly below the level 1 granularity. The next granularity is subdivided into twelve value groups (which we will identify with level 2 granularity), i.e., self-direction, power, security, conformity, benevolence, universalism, stimulation, hedonism, achievement, face, tradition, and humility. The finest value groups (which we will denote as level 3 granularity from now on) are the individual leave classes which are described below. It is worthwhile to mention that some classes are defined as both level 2 granularity as well as in level 3 granularity:

Openness to change:

- Self-direction – thought: Behind this argument is the wish to develop own ideas, the aim to know more and to discover.
- Self-direction – action: Here the argument expresses an intention of action, rather than a conviction or thought.
- Hedonism: here the person aims to have a good time, tries to enjoy life and tries to take advantage of fun opportunities.
- Stimulation: Here the intention is to focus on the novelty and risk aspects of behaviours and thoughts. In short, this value describes anything that stimulates the senses.

Self enhancement:

- Hedonism: (which can be found as well in the category Openness to change)

- Achievement: achievement, which is perceived within the social standards, and according to the rules of engagement. This value focuses on performance rather than materialistic matters.

- Power - dominance: This is value describes having power over people, which can be achieved by being the most influential compared to others, or to be the individual which determines the direction of things happening.

- Power – resources: Describes having power materialistically or socially which comes in terms of money, wealth, or a decent social status.

- Face: In that case the meaning behind the argument is that somebody does not want to be shamed by others or wants to protect one's public image. This person also wants to be treated with respect, honour, and dignity.

Conservation:

- Face: (which can be found as well in the category Self-Enhancement)

- Security – personal: Here, the important points are that somebody wants to have personal security and safety through a secure environment, a secure income as well as to be healthy.

- Security – societal: This value focuses on safety and stability in a wider society, rather than individual security.

- Tradition: Tries to maintain and preserve the cultural, family or religious traditions.

- Conformity – rules: The argument stresses on conforming a given set of rules, laws or formal obligations.

- Conformity – interpersonal: Which tries to avoid upsetting or harming other people.

- Humility: Here the main intention is to not draw attention, to be humble as well as to be happy with the given situation and to be happy with the things somebody has.

Self-transcendence:

- Humility: (which can be found as well in the category Conservation)

- Benevolence – caring: Tries to help, care and to be responsive to somebody close.

- Benevolence – dependability: Here the argument describes being a reliable and trustworthy member of a group of close people that those people have confidence in helping.

- Universalism – concern: Here the personal motivation is to protect vulnerable people, to give everybody an equal opportunity and to treat everyone duly.

- Universalism – nature: Stresses about the preservation of the environment.

- Universalism – tolerance: Here the argument tries to understand other people's views, and it wants to ensure peace and harmony even with arguments one strongly disagrees with.

2 Exploratory Data Analysis

After defining the classes and all levels of granularity the arguments are labelled with, another pivotal task prior to evaluating any LLM is to get a profound understanding of the structure of the dataset as well as being aware of any unknown inconsistencies (such as duplicate arguments with different labels or arguments with no labels at all) which could skew the results of the experiments. To do so, we have conducted an exploratory data analysis (EDA) which is implemented in `EDA.ipynb` of the public GitHub repository of this paper.

2.1 Initial insights

The first insight can be gained by counting the entries of each dataset from (Mirzakhmedova et al., 2023) which indeed has 8865 arguments, which are split into 5393 training arguments, 1896 validation arguments and 1576 test arguments. It is important to mention here that LLMs or any other smaller language models which are fine-tuned for this task, are not supposed to grasp any information about the validation- nor the test sets before the training phase of the model has been concluded. Otherwise, a data leakage might arise, which gives the model a wrong perfect score in the final evaluation. For this reason, only the training arguments have been considered.

The next step which has been done was to find any null (i.e. missing) values inside the data frame. As the Touché23 is a curated dataset no missing values have been found in the dataset as expected.

2.2 Feature analysis

Another vital aspect to analyse the given data is indubitably the connection between labels and its features. In the set of data, a column ’Premise’ occurs which embodies the required features to this multi-label classification task. Furthermore, each entry has exactly 20 columns where a zero denotes that the label is not associated with the respective premise and a one otherwise, in a one-hot encoding manner. To make this analysis more indicative, the given features have been pre-processed. More concretely, all letters have been transformed to lowercase characters and hashtags, punctuation, leading- and trailing white-spaces were removed. Further, multiple consecutive white-spaces have been replaced with a single space. This ensures that all the words are in a uniform format for further analysis. The first two premise properties which can be analysed is the distribution of length word- and character wise which has been visualized in figure 1 and figure 2:

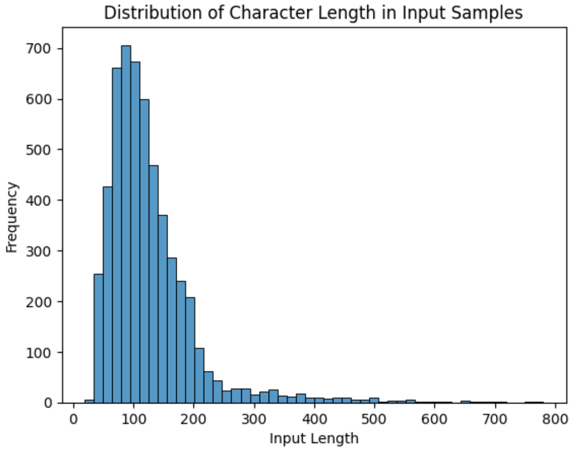

Figure 1: Character count

It has been discovered that the average premise length is 126 characters or 21 words. This minimum length has been found to be 19 characters or 3 words. On the other hand, the maximum length was found to be 780 characters or 133 words. By taking a closer look at the premises with only three words or less (which can be found in table 4), an inconsistency has been discovered by coincidence, that must be considered in the subsequent analysis of any LLM. The arguments with the ID E04046 and E07262 have an identical premise, however E07262 has the additional label ’Conformity: Benevolence’.

An additional aspect worthy of analysis involves an examination of the most frequently occurring words within this dataset. By solely counting the occurrence of every individual word and ordering them by occurrence one would obtain the following list:

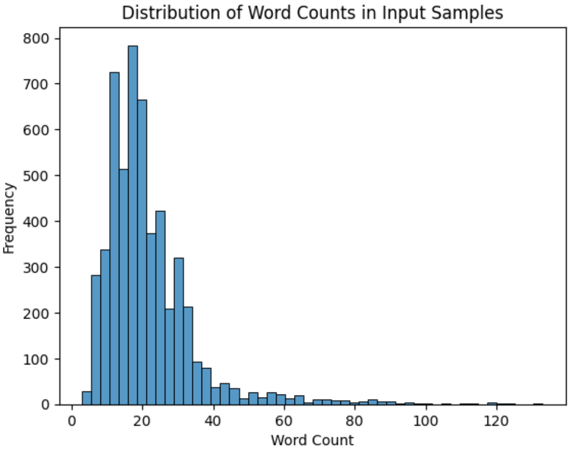

Figure 2: Word count

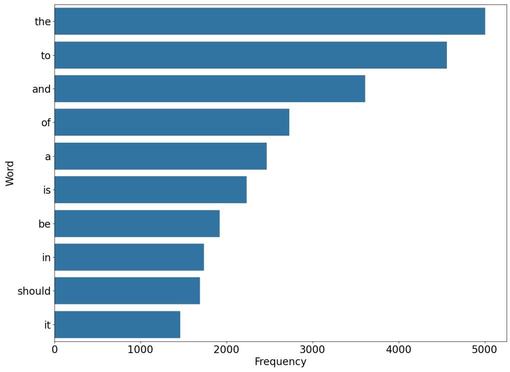

Figure 3: Most frequent words with no filtering (trimmed version)

A longer version of this graph can be found in the appendix in figure 14 due to its size. As one can grasp, the words with the highest quantity are ’the’, ’to’, ’and’, ’of’, ’a’, ’is’, ’be’, ’in’, ’should’ and ’it’. This of course is not a useful insight as most of those words are so-called ’stop words’. To make the output more meaningful one must ignore the stop words, which can be done by excluding words from this statistic which are contained in the ’scikit-learn’s stop words module’. By taking this into account one obtains the subsequent list, which again is a trimmed version. A more detailed version has been appended again as figure 15:

Here are the words with the highest quantity in

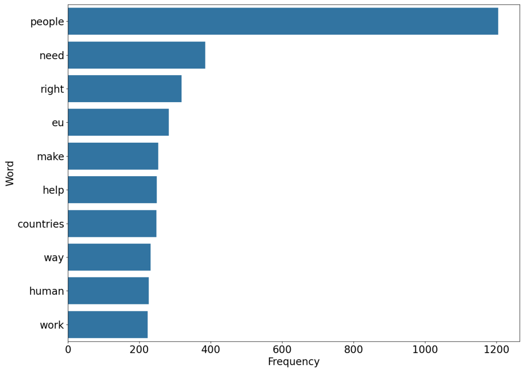

Figure 4: Most frequent words with with filtering scikit-learn's stop words out (trimmed version)

descending order: 'people', 'need', 'right', 'eu', 'make', 'help', 'countries', 'way', 'human' and 'work'. This ordering gives a better understanding as the previous one.
The next thing one can do is of course to filter the ordering of words by label, which can be done similarly. One only has to count all words which are in premises associated with this respective label. For the labels 'Conformity: interpersonal', 'Universalism: nature' for example, the most frequent words associated are 'people', 'telemarketing', 'language' and 'animals', 'zoos', 'human'. They are displayed in the figure 5 and figure 6 below. A more detailed version has also been added to the appendix in figure 16 and figure 17:

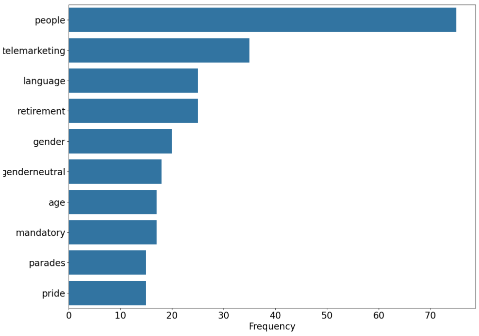

Figure 5: Most frequent words in 'Conformity: interpersonal' (filtered, trimmed version)

2.3 Label assignment distribution analysis

This chapter investigates the minimum-, maximum- and average number of labels assigned to each argument as well as its distribution. Further, this chapter reports the number of samples that were

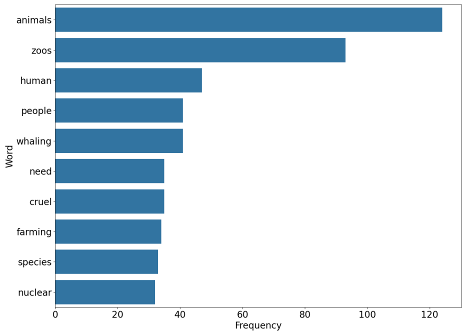

Figure 6: Most frequent words in 'Universalism: nature' (filtered, trimmed version)

assigned a certain label. This subsection will end with finding a correlation between the labels using the bi-variant analysis as well as the cosine similarity.

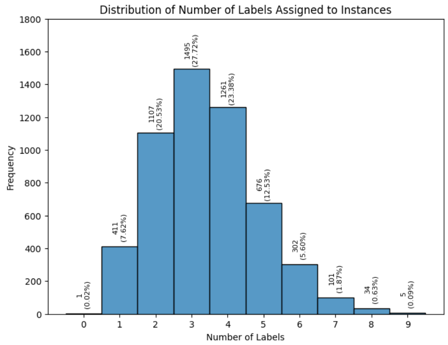

Figure 7: Distribution number of labels

As one can grasp from figure 7 there are three labels assigned per argument in the training data, but with a maximum of nine labels. An interesting fact about the analysed dataset that was discovered by doing this analysis was that the minimum number of labels assigned to an argument is zero, which means that this argument has not been classified at all. Hence, there is another inconsistency which must be considered when evaluating and training LLMs with this dataset.
Furthermore, one can gain information about the quantity of the labels in the dataset by counting the number of ones for each label-column and divide that number by the number of entries in the dataset. For this analysis all three data sets have been evaluated. This has been done and visualised in figure 8:
By looking closely in the bar diagram above, one could have noticed that the sum of all percentages of one colour does not add up to 100 percent. This

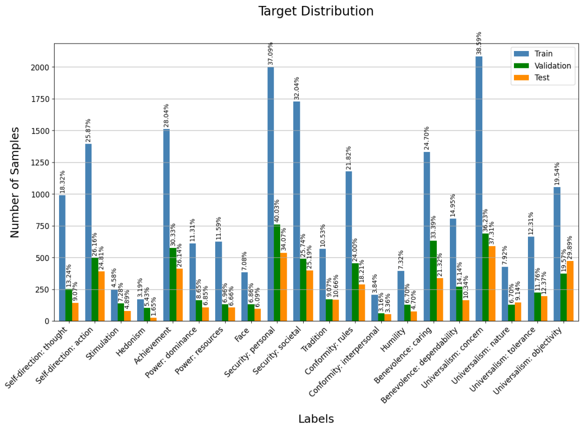

Figure 8: Distribution of labels in all data sets

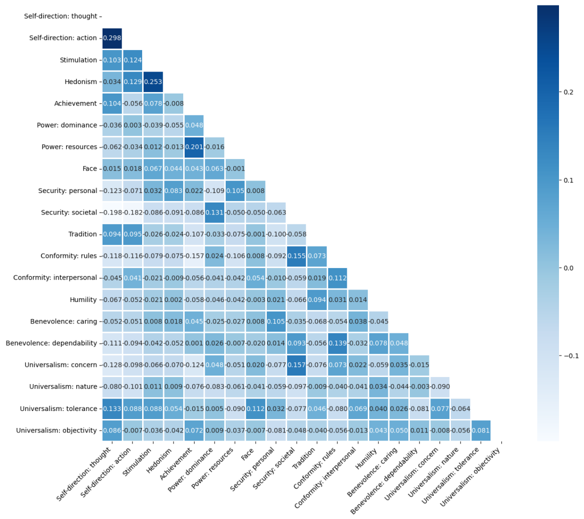

Figure 9: Bi-variant Analysis

is indeed the correct distribution as most premises in the dataset has more than one label assigned.

Another insight from figure 8 is that that the counts and ratio of labelled samples in all sets do not necessarily follow the same distribution, which are sometimes off by a quite noticeable margin. For instance, the label 'Security: societal' is assigned to 32.09% of the premises in the training dataset, whereas in the validation- and test dataset the percentage is close to 25%. Another example is the label 'Benevolence: caring' which is assigned to 24.70% of the arguments in the training dataset, but in the validation dataset it is assigned to 33.39%- and in the test dataset to 21.32% of the premises. This finding is helpful to know before doing any model training or fine-tuning. In an ideal setup, we would like the validation and test sets both to be representative of the training set. Otherwise, validation and test scores cannot be considered as robust and reliable indicators for the model's performance in real-world scenarios and the model might be biased towards the class which is represented most often.

The last insight might be desirable before starting the analysis of any LLM with this dataset, is how much the labels correlate by training data. This can be done by computing the correlation (in this paper the Spearman correlation formula has been used) and to visualise the given correlation factors in a heat-map, which has been done in figure 9: By looking at the figure 9, one can see that the correlation between the individual labels in training data is quite low. The highest correlated labels are 'Self-direction: thought' and 'Self-direction: action' with a correlation coefficient of 0.298. This means that if a premise is labelled 'Self-direction: thought' it is most likely that it is also labelled with 'Self-direction: action'. However, it is important to mention that a correlation of roughly 0.3 is not considered high by any means as high. This suggests that while there is a relationship between these two labels, there are likely other factors at play. It's also crucial to remember that correlation does not imply causation. So, even if these labels appear together, one does not necessarily cause the other.

Another handy tool to find correlation between labels is the cosine similarity analysis, which is scanning the dataset for semantically similar labels. Cosine similarity measures the cosine of the angle between two vectors in a multi-dimensional space (which is the vector representation of words in this case). This can be used to determine how similar two pieces of text are, irrespective of their size. By looking at figure 10 one can grasp that most of the entries are 0, which means that those labels are neither completely similar nor completely dissimilar. The highest correlation can be found between the labels 'Self-direction: action' and 'Self-direction: thought' with a correlation coefficient of 0.667. One can also see that there is also a correlation between each label of the group 'Universalism' with a coefficient of 0.5 each. This stands for a moderate degree of similarity between the two labels.

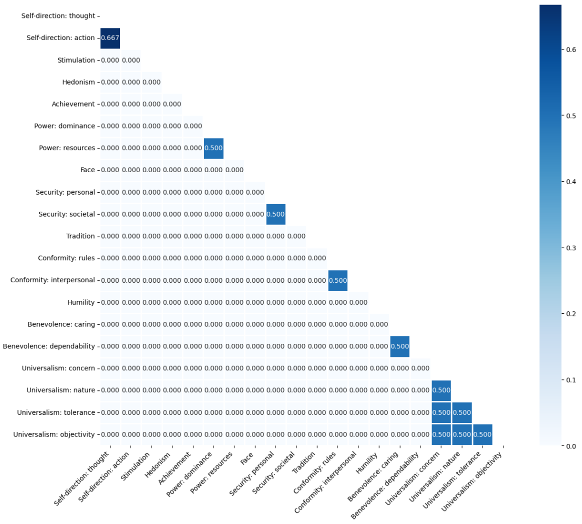

Figure 10: Cosine similarity between labels

2.4 Exploring Dataset Inconsistencies Further in Depth

This chapter aims to explore inconsistencies identified in the preceding chapter, specifically the presence of unlabelled premises and the existence of duplicate premises with disparate labels. For this, we have defined a function with finds duplicate values in each column of a given data frame, which will further store the respective details in a dictionary. In doing so, 80 entries (with pre-processed premises 89 entries) have been found in the training data which are duplicates. The same procedure can be done for the validation dataset, which had in total 34 duplicates (and 35 after pre-processing the premises), and for the test dataset which had 28 duplicates (and 32 with pre-processed entries). One might notice the slim difference between the number of duplicates found in the normal and the pre-processed data frame. This is because for a duplicate being recognised by the program, the characters must be identical. This means that a sentence with the word "National" will be labelled as a non-duplicate by any program following the same logic than the same sentence with the word "national". As an indisputable number of duplicates (156 out of 8865 samples) we also checked if there are any differences in "stance" and "conclusion" columns for these duplicated premise samples, to make sure that they are not intentionally duplicated to convey another meaningful human value. We found out that there are no differences in "stance" and "conclusion" columns for duplicated premise samples, except for one sample from validation data, which was only a difference in letter capitalization.

3 ChatGPT 3.5

The first large language model which was investigated in this paper was ChatGPT with the version 3.5, whose history and future is described in great detail in (Wu et al., 2023). This general LLM was released by OpenAI in the end of 2022 and was updated to ChatGPT 4 (and respectively the free version ChatGPT 3.5). It has surpassed the 100 million active users per month at the time of research. This LLM interacts in a chat window with the respective user, but it can also be used by API calls.

3.1 Setup

As ChatGPT 3.5 only provides an answer in response to a prompt the python package 'g4f'[3] has been used to submit messages to ChatGPT3.5 and then to retrieve the response to a txt-file. Because ChatGPT 3.5 is a general LLM which has not been fine tuned for this particular task, the task had to be re-introduced in every single prompt. This is the reason why every request sent to ChatGPT 3.5 had the shape of `<initial prompt\n\n` 10×`triplet>`, where the set of triplets is different for each prompt. Every prompt consisted of 10 triplets to make the conversation with ChatGPT more efficient. For more than ten triplets per prompt ChatGPT had problems processing inputs as such, as it started to provide more answers than the number of triplets. Further it did start to repeat its last answer many times. Each triplet has a justification-, a stance- as well as a premise entry. The triplet can also be imagined as a single row in the test dataset. On the other hand, `initial prompt` is re-introducing the task to the LLM. This includes a role which the user is in - in this case a psychologist and a scientist. Then the format of the given triplet, the desired output (format) and the classes are described. For the experiments three different initial prompts have been written per role. The first type of initial prompt provides three different examples to the LLM. The second type contains one example of each class. Lastly, the third initial prompt didn't contain any examples. This task is done by `pipeline.ipynb` which is included in the public GitHub repository.
The next step was to further pre-process the labels provided by ChatGPT 3.5, which unfortunately

[3] GitHub: https://github.com/xtekky/gpt4free

could not be done fully automatically. To make the manual pre-processing of the raw data manageable `prepare_gpt_results.ipynb` was written. This script starts by removing all occurrences of ’*’ as well as all lines which start with ’(")’ or which do not start with ’(’ at all, which all denote an invalid answer given by the LLM, which were roughly ten percent of the returned answers. In some responses ChatGPT introduced new labels which were not part of the original set of labels. This is the reason why this script first marked all occurrences of invalid labels, which were manually interpreted and then automatically replaced with some fitting synonym of the respective labels. The number of the occurrences of the made up labels as well as the labels are provided in tables 1, 2 and 3 where the number of invalid labels can be found at the end of each table, where all invalid labels are underlined in the corresponding table. In some cases it was possible to re-run `pipeline.ipynb` for the prompts with an invalid output format, to obtain a valid answer.
The last step was to interpret the pre-processed data which is done by `analyze_results.ipynb`. Which will be covered by the subsequent subsection.

3.2 Results for finest granularity

This chapter will highlight the results of the experiments carried out, which have been analysed with the finest granularity. This has been introduced in the first chapter as granularity level 3, which consists of 20 value categories. In the following, the three previously described initial prompts will be discussed separately. To improve readability, the results have only been provided with the first 6 digits after the comma.
The first test case presented three different examples to ChatGPT 3.5 prior carrying out the labelling, where the results are summarized in table 1:

> - *Input: ("We should end the use of economic sanctions", "contra", "Economic sanctions provide security and ensure that citizens are treated fairly")*
>   *Your response: (societal, concern)*
>
> - *Input: ("We need a better migration policy", "pro", "Discussing what happened in the past between Africa and Europe is useless. All slaves and their owners died a long time ago. You cannot blame the grandchildren")*
>   *Your response: (concern)*
>
> - *Input: ("Rapists should be tortured", "contra", "Throughout India, many false rape cases are being registered these days. Torturing all of the accused persons causes torture to innocent persons too.")*
>   *Your response: (societal, concern)*
>
> (Mirzakhmedova et al., 2023)

For the psychologist version of the prompt ChatGPT3.5 returned 2717 labels which didn’t have to be removed due to an incorrect answering format (which will be referred to as ’answers of invalid format’ in the subsequent paragraphs). Here it is worth mentioning that the LLM hasn’t returned any invalid labels. That is because they were manually mapped to an equivalent valid label. From looking at the true- and false positives as well as at the true- and false negatives this yields a precision of 0.377622, a recall of 0.215049, an accuracy of 0.827538 and a specificity of 0.936782. From that the F1-score can be computed: 0.274038.
For the next test case only one word has been changed from the initial prompt: the role of the user was changed to a scientist, which made ChatGPT3.5 label some phrases differently than before. Here the LLM returned 2739 answers of valid format, where the labels of ten answers have been made up by the LLM, which equals to 0.37% of the answers of valid format: religion, equality, label_1.1, label_1.2 and freedom. Here, the precision was 0.374309, the recall was 0.213162, the accuracy was 0.826967 and the specificity was 0.936446. This yields a F1-score of around 0.271634.
The next test case presented exactly one example of each category to the LLM before the labelling task, where the results are summarized in table 2:

> - *Input: ("We should cancel pride parades", "in favor of", "pride parades create a huge disturbance")*
>   *Your response: (rules, interpersonal)*
>
> - *Input: ("We should ban the use of child actors", "in favor of", "child actors lose the sense of a proper childhood.")*
>   *Your response: (Stimulation, Hedonism, personal)*
>
> - *Input: ("We should adopt a multi-party system", "against", "multi-party systems slow down what gets done because we have too many different sides trying to come to an agreement")*
>   *Your response: (action, Achievement, dominance, objectivity)*

> - *Input: ("We should ban missionary work", "in favor of", "if we ban missionary work, then a lot less people would be seeing propaganda.")*
>   *Your response: (thought, caring, tolerance, objectivity)*
> - *Input: ("We should subsidize student loans", "against", "it isn't the obligation of any tax payer to give money to help someone else get an education. that is a personal choice, and if they want to go to college they need to pay for it on their own.")*
>   *Your response: (Achievement, resources, personal, dependability)*
> - *Input: ("We should end mandatory retirement", "in favor of", "mandatory retirement is purely age discrimination")*
>   *Your response: (action, concern)*
> - *Input: ("We should adopt atheism", "against", "Atheism is godless and fundamentally lacking in a coherent moral compass, therefore it should not be adopted.")*
>   *Your response: (Face, societal, Tradition, dependability)*
> - *Input: ("We should ban cosmetic surgery", "in favor of", "cosmetic surgery should be banned. god made you a certain way and these procedures are going against that. besides, it could be dangerous as joan rivers died from complications during surgery.")*
>   *Your response: (personal, Tradition, Humility, nature)*
>
> (Mirzakhmedova et al., 2023)

For the psychologist version of the prompt the LLM returned 3224 answers in of valid format, where 17 triplets have been labelled with classes ChatGPT3.5 has made up, which equals to 0.53% of the answers. The non-existent classes were: autonomy, equality, ethics, cultural, labelling, family, justice, freedom of speech, employment and freedom. With the given test phrases this version of the prompt reached a precision of 0.370896, a recall of 0.246279, an accuracy of 0.822684 and a specificity of 0.925492. Considering all that, one can compute the F1-score, which is 0.296007.

Now the role of the user in the prompt was introduced as a scientist again. Here the LLM returned 3216 answers of valid format, where 19 answers were not part of the given set of labels, which is 0.59% of the labels: gender, autonomy, freedom of speech, free speech, health, ethics, freedom, privacy, religious and equality. Here the case had a precision of 0.368870, a recall of 0.244393, an accuracy of 0.822335 and a specificity of about 0.925417. This gives a F1-score of around 0.293998.

The third initial prompt didn't contain any examples, which means that the LLM started to label the given triplets without looking at any triplets which have been manually labelled. The described results are summarized in table 3.

With the psychologist version of the prompt, 3203 answers of valid format have been returned. From those 3203 answers 39 prompts have been labelled with a label made up by the LLM, which equals to 1.22% of the responses of valid format. Those labels were: respect, jealousy, promote, support, comply, improve, natural, discourage, human rights, free choice, justice, fairness, free speech, equality and freedom. In this test case the computed precision was 0.321428, the recall 0.211276, the accuracy of 0.813102 and a specificity of 0.920445. This equals to a F1-score of 0.254963.

When changing the role of the user to a scientist again and keeping the same initial prompt with this minor change the LLM returned 3160 answers of valid format, where 32 triplets were labelled with made up labels, which equals to 1.01%. The labels made up by the LLM were key made up by ChatGPT: freedom, fairness, equality, freedom of speech, law, constitutional right, gender equality, family, justice, health, security, patriotism, human rights, free expression, privacy and humanity. The carried-out experiments showed a precision of 0.333872, a recall of 0.216516, an accuracy of 0.816021 and a specificity of 0.922950. This combined yielded a F1-score of 0.262682.

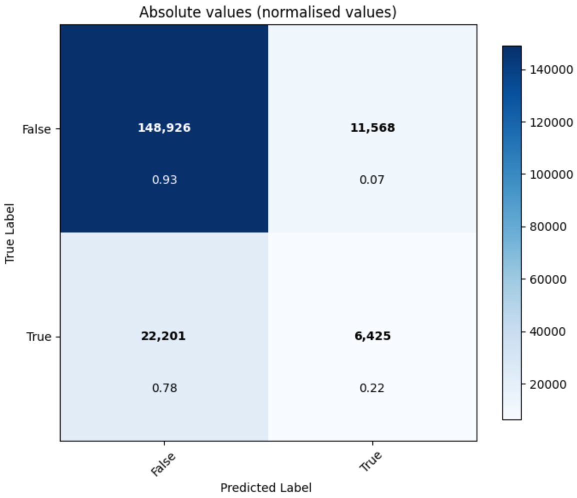

Figure 11: Level 3: Prompts 1 to 3 accumulated for both roles

To summarise the previously gathered information: Looking at the average F1-score (≈ 0.275553) which ChatGPT3.5 obtained for labelling the given arguments with the introduced labels, it becomes fairly evident that ChatGPT3.5 could only poorly predict the labels to the given human values. This gets very clear when looking at figure 11, which is the confusion matrix for granularity level three using the accumulated number of true- and false positives as well as the true- and false negatives. One can see that only 22.44% labels actually have been correctly predicted. On the other hand, one can see that ChatGPT could correctly predict in roughly 92.79% of the cases, that a given argument does not belong to a certain class.

In the subsequent chapter, the labels will be generalized by grouping them according to level two and level one granularity which has been covered by chapter one.

3.3 Results for a coarser granularity

This chapter will first start by summarising the results for granularity level two, then continue with granularity level one and then compare the results with the ones of the previous subsection of this paper.

To make both results comparable the raw data of the previous chapter have been re-used, but in the analysed results script the labels have been summarised into enumeration sets, and a label has been predicted correctly, if the predicted label and the true label are in the same enumeration set.

To improve the readability of the text, the calculated numbers will only be provided to the first 6 digits.

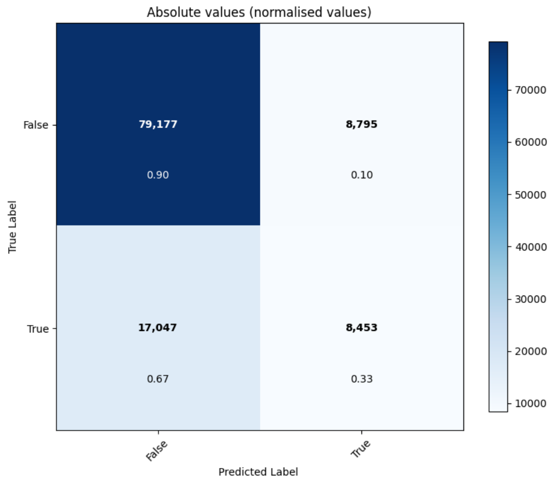

Figure 12: Level 2: Prompts 1 to 3 accumulated for both roles

The first chapter shows that granularity level reduces the 20 labels to twelve. However, when inspecting figure 12 one can hardly see the difference to figure 11. At second glance, one can see that both the number of false-positive labels as well as the number of false-negatives have been significantly reduced.

With the given number true- and false positive as well as the true- and false negative labels one can compute the five classification numbers again. Averaging the results for each test case for this granularity, the precision was 0.490085, the recall 0.331490, the accuracy 0.772260 and the specificity 0.900025. Combining this information, the overall F1-score is 0.395480, which represents a significant improvement to the previous F1-score.

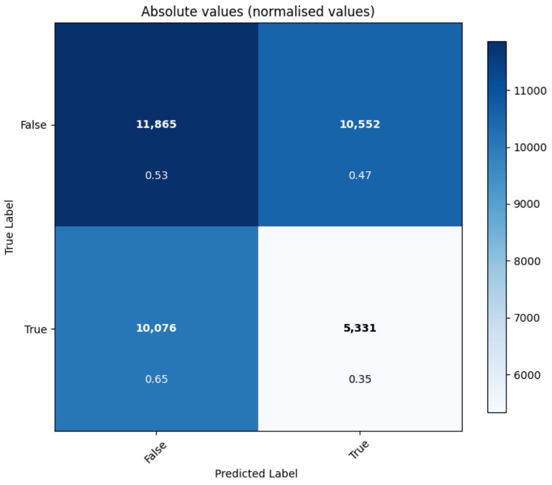

Figure 13: Level 1: Prompts 1 to 3 accumulated for both roles

The next granularity, which will be covered by this paper, is granularity level 1, which groups the labels into four groups.

By looking at figure 13 one can see that having such a course granularity has immensely improved the classifying performance of the large language mode. Both false-positives and false-negatives have dropped again and apart from the false positives the number of true-positive, true-negative and the false negative is very close to 10500.

Using those four values again one can compute recall, precision, accuracy, specificity and the F1-score again. With this granularity ChatGPT3.5 achieved a precision of 0.664358, a recall of 0.511537, an accuracy of 0.592666 and a specificity of 0.689986, which leads to a F1-score of 0.799241.

4 Llama2-7b

This chapter covers the results from the LLM Llama2-7b from Meta.(lla) For this large language models the experiments only covered the test cases with the finest granularity, i.e., level 3 granularity with 20 value categories. For this series of experiments conducted Hugging Face has been used to pass the prompts to hugging face and to retrieve the answer, which has been done in `get_LLaMa_2_values.ipynb` of the public dataset which is referred to in page one. The structure of the prompt was identical to the one described in chapter 3.1 as well as chapter 3.2. Due to the fixed input format of the Hugging Face API a pre- and post-fix had to be added such that the final prompt passed to the LLM had the following structure: `[INST]<>{{ <PROMPT> }}<>{{user_prompt}}[/INST]`. As the the prompts from Llama2-7b are identical the the one from ChatGPT3.5 - apart from this minor change - the same analyzing script (`analyze_results.ipynb`) has been used.

4.1 Results

As mentioned in the short introduction of this chapter the only granularity which has been considered was the level3 granularity (with 20 different labels). To make this section readable, only the first six digits will be provided by this paper.

The first prompt (which has been introduced in chapter 3.2) passed three examples to Llama2-7b. For the psychologist version of the prompt this yielded a precision of 0.289419 a recall of 0.108857, a accuracy of 0.827522, a specificity of 0.953245 and a F1-score of 0.158208. For the scientist version of the prompt a precision of 0.297423, a recall of 0.099738, an accuracy of 0.830772, a specificity of 0.958753 and a F1-score of 0.149382.

The second prompt (which can also be found in chapter 3.2) had one representative sample per label. For the psychologist version of the prompt a precision of 0.182894, a recall of 0.040930, an accuracy of 0.830210, a specificity of 0.968062 and a F1-score of 0.066891 have been obtained. On the other hand, the scientist version of the prompt had a precision of 0.171951, a recall of 0.040331, an accuracy of 0.828340, a specificity of 0.966056 and a F1 score of 0.065338.

For the last version of the prompt no examples at all have been passed to Llama2-7b. The psychologist version of the prompt received a precision of 0.156389, a recall of 0.052854, an accuracy of 0.817857 a specificity of 0.950547 and a F1-score of 0.079006. The scientist version of the prompt had a precision of 0.138912, a recall of 0.046158, an accuracy of 0.815328, a specificity of 0.949929 and a F1-score of 0.069292.

5 Conclusion

In this paper Touché's dataset was examined in a meticulous manner. First some overall insights were created such as the number of elements in the given dataset, and how many of those elements were used for training, validation and testing.

Then it has been discovered that the average argument length has been 21 words (or roughly 126 characters) and that the average argument was labelled with three labels. Furthermore, it has been ascertained that the arguments which are associated with a certain label have set of keywords which they are associated with. For instance, the label 'Conformity: interpersonal' that most frequent words have been 'people', 'telemarketing', 'language'. For 'Universalism: nature' they have been 'animals', 'zoos', 'human'.

By conducting a more thorough analysis of the dataset, it has been determined that the counts and the ratio of labelled samples in all sets do not follow the same distribution, and are in fact off by a great margin, which might cause any LLM which is fine-tuned with this dataset being biased.

It also has been discovered that various inconsistencies have been found throughout the given dataset, including in the training, validation, and testing sets. Some of the given arguments have not been labelled at all, where various arguments re-appeared in the dataset with the same premise as well as the same stance, but different labels assigned which may be a problem. This might have skewed the overall score, like the F1-score, of the trained LLMs for that dataset which have been listed in (Kiesel et al., 2023) on page 6.

This paper also conducted experiments with ChatGPT3.5 and Llama2-7b which included using the dataset from (Mirzakhmedova et al., 2023) and letting general LLMs to label the given arguments. In summary, the results show that role-based prompting produced similar levels of accuracy for both the scientist and psychologist role models. However, when comparing the F1 score of ChatGPT3.5 (averaging 0.28) to the F1 score of the fine-tuned LLMs in Touché's paper (approximately 0.56 for the top 5 LLMs), there is a noticeable difference that the general LLM performed poorly. By presenting one labelled argument per label type, the F1 score improved from 0.27 to 0.29 for ChatGPT3.5. On the other hand, by not giving any exemplary arguments to ChatGPT3.5 the average F1 score has been at around 0.25. Finally, it was found that by grouping similar labels together to form label groups with a coarser granularity, the F1 score marginally improved to 0.395480 with granularity level 2, which has 12 label groups, and granularity level 1, which has 4 value groups, with an F1 score of around 0.8. For Llama2-7b only experiment have been done to the finest granularity which had 20 different labels. However, the F1-score was even worse than for ChatGPT3.5 with an average F1-score of roughly 0.098019. Which was very interesting to see when conducting those experiments was that the prompt with three exemplary labelled prompts performed better (about 0.153795) than the version which had one labelled prompt per label class (about 0.066114).

References

Llama2-7b. https://ai.meta.com/llama/. Accessed: January 25, 2024.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. Human value detection 2024. https://touche.webis.de/clef24/touche24-web/human-value-detection.html#task. Accessed Dec 1, 2023.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. 2023. Semeval-2023 task 4: Valueeval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval'23)*, pages 2287–2303. Association for Computational Linguistics.

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke Xiaoni Cai, Valentin Barriere, Doratossadat Dastgheib, Omid Ghahroodi, Mohammad Ali Sadraei Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2023. The Touché23-ValueEval Dataset for Identifying Human Values behind Arguments. *CoRR*, abs/2301.13771.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.

6 Appendix

Psychologist Prompt		Scientist Prompt	
Label	**Absolute (relative) Occurrence**	**Label**	**Absolute (relative) Occurrence**
hedonism	5 (0.0018)	religion	1 (0.0004)
stimulation	8 (0.0029)	equality	1 (0.0004)
face	9 (0.0033)	label_1.1	1 (0.0004)
tolerance	12 (0.0044)	label_1.2	1 (0.0004)
objectivity	18 (0.0066)	hedonism	2 (0.0007)
dependability	21 (0.0077)	face	5 (0.0018)
humility	24 (0.0088)	freedom	6 (0.0022)
dominance	27 (0.0099)	stimulation	10 (0.0037)
caring	31 0.0114)	dominance	14 (0.0051)
achievement	47 (0.0173)	thought	15 (0.0055)
tradition	50 (0.0184)	caring	17 (0.0062)
nature	57 (0.0210)	dependability	19 (0.0069)
action	71 (0.0261)	humility	20 (0.0073)
thought	73 (0.0269)	tolerance	21 (0.0077)
interpersonal	86 (0.0317)	objectivity	2 4(0.0088)
rules	87 (0.0320)	achievement	32 (0.0117)
personal	190 (0.0699)	tradition	39(0.0142)
resources	216 (0.0795)	nature	52 (0.0190)
concern	826 (0.3040)	action	55 (0.0201)
society	859 (0.3162)	rules	64 (0.0234)
		interpersonal	79(0.0288)
		resources	149 (0.0544)
		personal	165 (0.0602)
		society	973 (0.3552)
		concern	974 (0.3556)
Made up:	0 (0.00%)	**Made up**:	10 (0.37%)
Total: 2717		**Total**: 2739	

Table 1: ChatGPT3.5: Results prompt 1

Psychologist Prompt		Scientist Prompt	
Label	**Absolute (relative) Occurrence**	**Label**	**Absolute (relative) Occurrence**
autonomy	1 (0.0003)	gender	1 (0.0003)
equality	1 (0.0003)	autonomy	1 (0.0003)
ethics	1 (0.0003)	freedom of speech	1 (0.0003)
cultural	1 (0.0003)	free speech	1 (0.0003)
labeling	1 (0.0003)	health	1 (0.0003)
family	1 (0.0003)	ethics	1 (0.0003)
justice	2 (0.0006)	freedom	2 (0.0006)
freedom of speech	2 (0.0006)	privacy	2 (0.0006)
hedonism	2 (0.0006)	religious	3 (0.0009)
employment	2 (0.0006)	hedonism	5 (0.0016)
freedom	5 (0.0016)	equality	6 (0.0019)
face	16 (0.0050)	humility	17 (0.0053)
humility	21 (0.0065)	stimulation	20 (0.0062)
stimulation	25 (0.0078)	face	23 (0.0072)
dominance	32 (0.0099)	dominance	34 (0.0106)
tolerance	43 (0.0133)	tradition	46 (0.0143)
tradition	48 (0.0149)	tolerance	47 (0.0146)
achievement	74 (0.0230)	nature	73 (0.0227)
caring	80 (0.0248)	caring	88 (0.0274)
nature	82 (0.0254)	achievement	96 (0.0299)
interpersonal	106 (0.0329)	interpersonal	102 (0.0317)
dependability	118 (0.0366)	dependability	142 (0.0442)
thought	208 (0.0645)	thought	178 (0.0553)
rules	225 (0.0698)	rules	211 (0.0656)
society	249 (0.0772)	personal	215 (0.0669)
personal	252 (0.0782)	society	261 (0.0812)
objectivity	253 (0.0785)	objectivity	284 (0.0883)
resources	303 (0.0940)	resources	291 (0.0905)
action	465 (0.1442)	action	472 (0.1468)
concern	605 (0.1877)	concern	592 (0.1841)
Made up:	17 (0.53%)	**Made up**:	19 (0.59%)
Total: 3224		**Total**: 3216	

Table 2: ChatGPT3.5: Results prompt 2

Psychologist Prompt		Scientist Prompt	
Label	**Absolute (relative) Occurrence**	**Label**	**Absolute (relative) Occurrence**
respect	1 (0.0003)	law	1 (0.0003)
jealousy	1 (0.0003)	constitutional right	1 (0.0003)
promote	1 (0.0003)	family	1 (0.0003)
support	1 (0.0003)	security	1 (0.0003)
comply	1 (0.0003)	patriotism	1 (0.0003)
improve	1 (0.0003)	human rights	1 (0.0003)
natural	1 (0.0003)	free expression	1 (0.0003)
discourage	1 (0.0003)	privacy	1 (0.0003)
human rights	1 (0.0003)	humanity	1 (0.0003)
free choice	2 (0.0006)	freedom of speech	2 (0.0006)
justice	2 (0.0006)	justice	2 (0.0006)
fairness	3 (0.0009)	health	2 (0.0006)
free speech	3 (0.0009)	fairness	3 (0.0009)
equality	9 (0.0028)	gender equality	3 (0.0009)
freedom	11 (0.0034)	freedom	4 (0.0013)
hedonism	16 (0.0050)	equality	7 (0.0022)
face	17 (0.0053)	hedonism	16 (0.0051)
tolerance	24 (0.0075)	tolerance	21 (0.0066)
humility	26 (0.0081)	stimulation	24 (0.0076)
objectivity	27 (0.0084)	face	25 (0.0079)
stimulation	28 (0.0087)	humility	29 (0.0092)
dependability	38 (0.0119)	objectivity	31 (0.0098)
dominance	46 (0.0144)	dependability	37 (0.0117)
nature	67 (0.0209)	dominance	61 (0.0193)
tradition	75 (0.0234)	nature	68 (0.0215)
caring	89 (0.0278)	tradition	91 (0.0288)
interpersonal	99 (0.0309)	achievement	103 (0.0326)
achievement	117 (0.0365)	caring	113 (0.0358)
action	207 (0.0646)	interpersonal	117 (0.0370)
rules	259 (0.0809)	action	152 (0.0481)
concern	271 (0.0846)	concern	272 (0.0861)
resources	338 (0.1055)	thought	294 (0.0930)
thought	405 (0.1264)	rules	307 (0.0972)
personal	493 (0.1539)	resources	310 (0.0981)
society	522 (0.1630)	personal	451 (0.1427)
		society	606 (0.1918)
Made up:	39 (1.22%)	**Made up**:	32 (1.01%)
Total: 3203		**Total**: 3160	

Table 3: ChatGPT3.5: Results prompt 3

line	argument id	premise	pre-processed premise	labels
4792	D27081	Failed biometric authentication.	failed biometric authentication	[Self-direction: action]
4951	E04046	Migrants sell drugs.	migrants sell drugs	[Security: societal, Conformity: rules]
5366	E07262	Migrants sell drugs.	migrants sell drugs	[Security: societal, Conformity: rules, Benevolence: caring, Benevolence: dependability]

Table 4: Inconsistency 1: Same premise, different labelling.

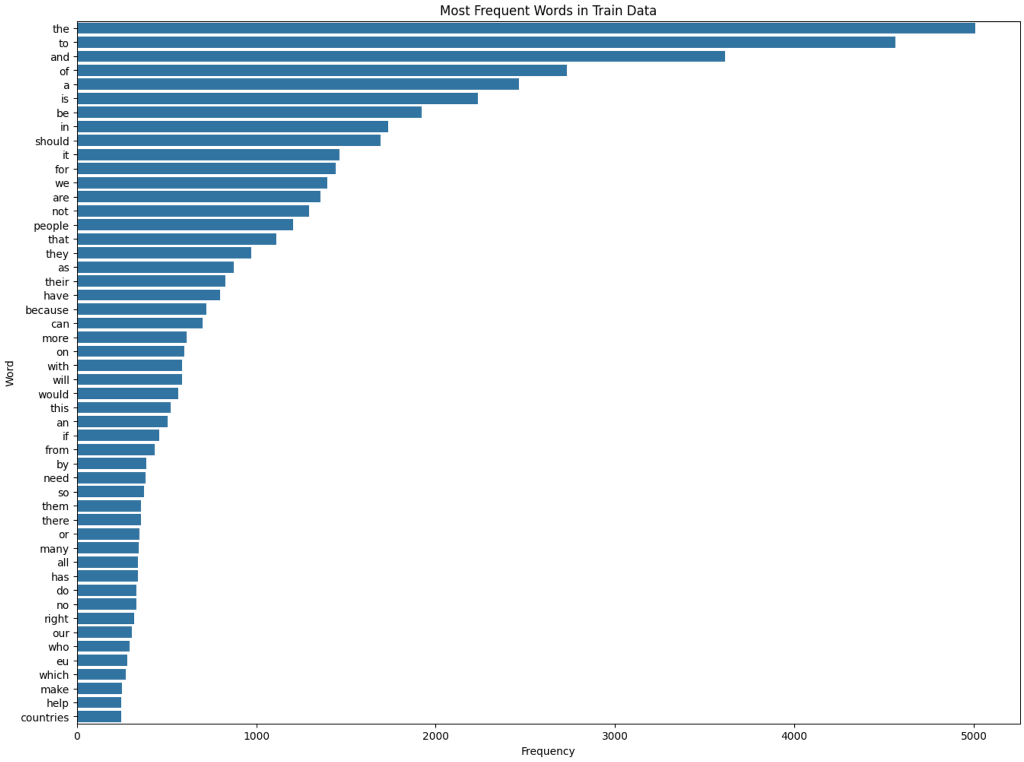

Figure 14: Most frequent words with no filtering

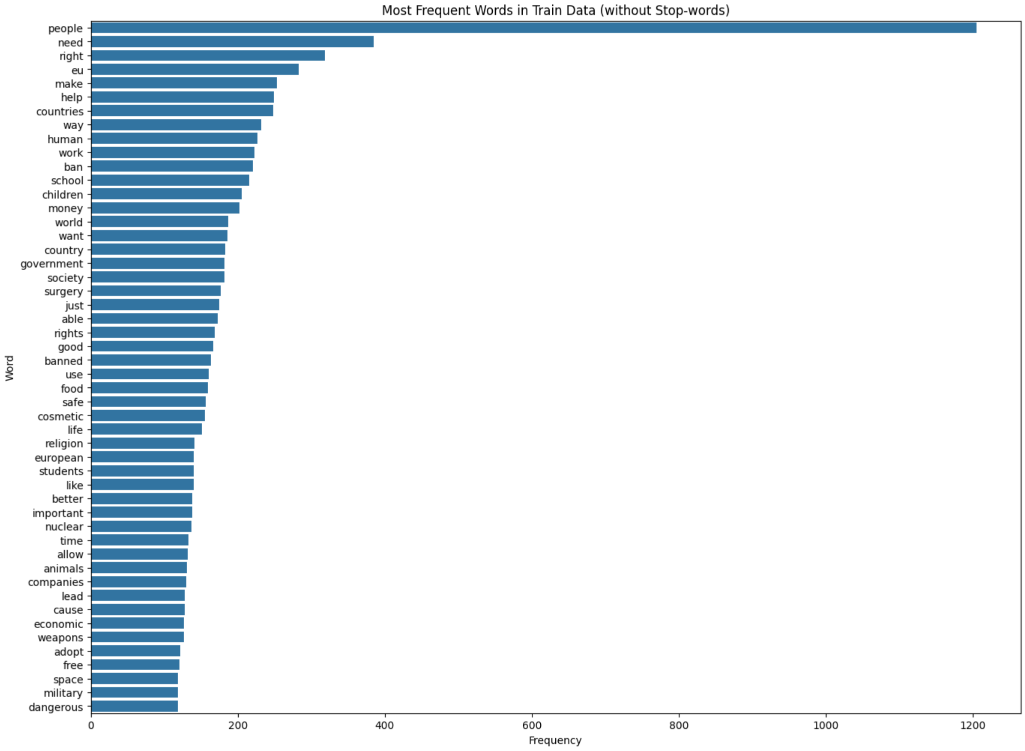

Figure 15: Most frequent words with with filtering scikit-learn’s stop words out

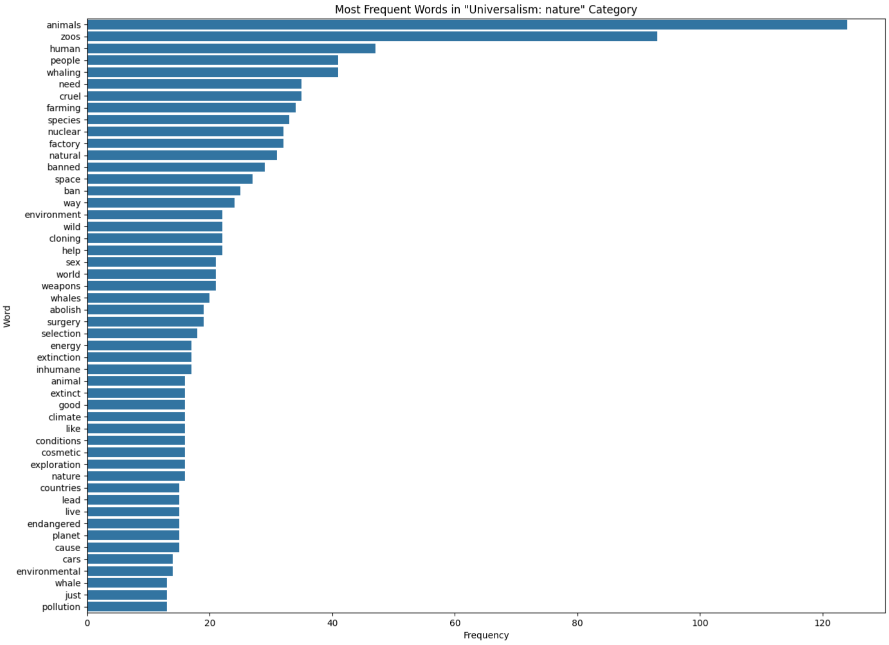

Figure 16: Most frequent words in ’Universalism: nature’ (filtered)

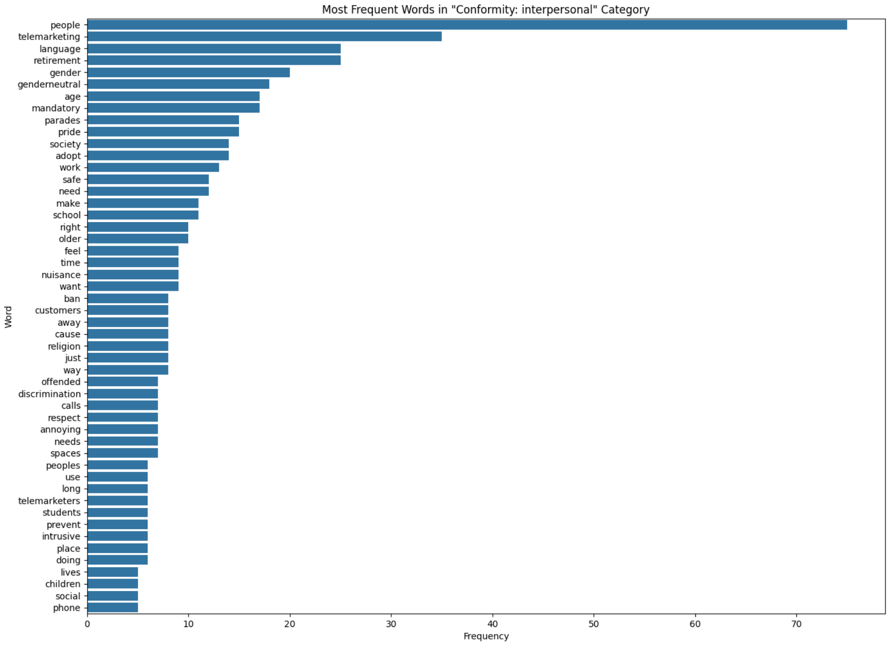

Figure 17: Most frequent words in ’Conformity: interper- sonal’ (filtered)