SINAI at PAN 2024 Oppositional Thinking Analysis: Exploring the Fine-Tuning Performance of Large Language Models

Notebook for PAN at CLEF 2024

María Estrella Vallecillo-Rodríguez[1], María Teresa Martín-Valdivia[1] and Arturo Montejo-Ráez[1]

[1] *Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071, Spain*

Abstract

This article describes the participation of the SINAI research group in the shared task "Oppositional Thinking Analysis: Conspiracy theories vs critical thinking narratives" in CLEF 2024. This task is composed of 2 subtasks subtask 1 which consists of a binary classification between critical and conspiracy texts and subtask 2 which consists of a token-level classification of the element of the oppositional narrative. The proposed system for both subtasks consists of the use of LLMs (LLaMA3 or GPT-3.5) where we apply an instruction tuned for the specific subtask. We think that these types of models have more knowledge and can reason to distinguish each type of text or elements of the texts and the instruction tuned will potentiate this, helping the models to distinguish between the classes. In the final leaderboard, our proposal obtained 3rd and 1st place for task 1 in English and Spanish respectively. In subtask 2 our systems reached the 18th position for English and 17th for Spanish.

Keywords

Large Language Models, QLoRA, Zero-Shot Learning, Oppositional Thinking Analysis,

1. Introduction

Nowadays, social networks are the most widely used means of communication by people. In them, users share various aspects of their lives, express their opinions, ideas, and even share current news. The problem is that not all the news published on social networks are true and users sometimes just by reading them are already spreading them on the network, without stopping to check the information. One type of message that is harmful to the social networking community is conspiracy theories, defined by the European Union [1] as: "The belief that certain events or situations are secretly manipulated behind the scenes by powerful forces with negative intentions". These theories are harmful because they can generate serious consequences in society, such as spreading distrust in public institutions or scientific information, feeding discrimination, and justifying hate crimes, among other consequences. However, there are other types of texts that can be found in social networks known as critical thinking narratives. In them, users express their opinions, sometimes argued and sometimes based on events that have happened to them or to acquaintances. A more concrete definition of what critical thinking narratives are is provided by the Oxford dictionary [2] "the process of analyzing information in order to make a logical decision about the extent to which you believe something to be true or false". If this critical thinking issues a judgment that opposes the main idea, we will be talking about an oppositional critical thinking narrative.

These two narratives explained above are challenging to distinguish, especially for language models that analyze social network content. Therefore, the organizers of the shared task "Oppositional Thinking

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

† These authors contributed equally.

mevallec@ujaen.es (M. E. Vallecillo-Rodríguez); maite@ujaen.es (M. T. Martín-Valdivia); amontejo@ujaen.es (A. Montejo-Ráez)

0000-0001-7140-6268 (M. E. Vallecillo-Rodríguez); 0000-0002-2874-0401 (M. T. Martín-Valdivia); 0000-0002-8643-2714 (A. Montejo-Ráez)

Analysis” in the PAN Lab [3] of CLEF 2024 propose 2 subtasks. The first subtask is to distinguish the conspiratorial narrative from other oppositional narratives that do not express a conspiratorial mindset (i.e., critical thinking). This is a binary classification task between two classes (CRITICAL or CONSPIRATIVE). The second subtask is to identify the key elements of a narrative that fuels intergroup conflict in oppositional thinking in online messages. This task is a token-level classification task in which models have to recognize text spaces corresponding to such key elements of oppositionalist narratives (AGENT, FACILITATOR, VICTIM, CAMPAIGNER, TARGET, NEGATIVE_EFFECT). These two subtasks are proposed to be applied in English and Spanish messages. The data of the task is extracted from the Telegram social network and is related to the COVID-19 pandemic.

Our proposal consists of the use of generative LLMs such as GPT-3.5 and LLaMA3-8B-instruct that are trained with instructions to detect conspiratorial and critical texts as well as the different elements of these narratives. To train the models we want to apply QLoRA, which is a method to train LLMs efficiently, and the OpenAI API. We think that this adaptation to the task will be crucial to help the models learn the differences between the different classes. With this proposal, we intend to study how the training of the models affects the achievement of the objectives of the proposed subtasks, and what differences exist between the size of the models, and their performance. There are recent studies that focus on the study of conspiracy theories in social networks, such as [4] in which the authors create a dataset that includes accounts dedicated to conspiracy theories and a control group of randomly selected users. They then perform a comparative analysis of the topics covered, profile characteristics, and behaviors. Using machine learning algorithms and features from bot, troll, and linguistic literature, they successfully classified conspiracy theory users with high accuracy. In contrast, other studies attempt to use and analyze the performance of generative LLMs to detect conspiratorial texts. Diab et al. [5] tries to address the detection of conspiracy theories by training a BERT model and then compare with the performance of the GPT model without applying any training to it. Their study finds that GPT fails to apply logical reasoning. However, other studies such as [6] which focuses on detecting conspiratorial Telegram messages in German, show a comparison between applying supervised tuning approaches (BERT models) and instruction-based approaches (LLaMA2, GPT-3.5 and GPT-4), which require little or no additional training data. Their work shows that both approaches can be used effectively, highlighting that among the highest results is GPT-4 with Zero-Shot Learning (ZSL) instruction and including a definition of what a conspiracy theory is. Peskine et al. [7] attempt to generate definitions from examples and use them for zero classification of fine-grained multi-label conspiracy theory. They show that improving class label definitions has a direct consequence on subsequent classification results. This makes us think that it is very important to refine the instruction we give to the model. Some studies analyze how well instruction-based models perform if they adjust a task. An example of such studies is [8] in which they use a LLaMA model containing emotional information and apply training based on different instructions (emotion recognition, sentiment, and conspiracy theories). Their results show that this model largely outperforms several open-source domain-general LLMs.

The remainder of the paper is organized as follows: Section 2 presents an overview of important details about the proposed system for the shared task. The used data and the methodology followed to achieve the goal of the task are described in Section 3. In Section 4 we show the results obtained in our experiments during the development phase and the evaluation phase. Finally, we conclude with a discussion in Section 5.

2. System overview

The developed system to achieve Oppositional Thinking Analysis shared task [9] at CLEF 2024 is described in this section.

To achieve both subtasks we want to study how LLMs such as LLaMA3 or GPT-3.5 which are generative models can be adapted to a classification task as proposed in this shared task. In addition we want to study whether the differences between the size of the models influenced the classification of each text. For this reason, we plan to apply an instruction-based training of the models. The first step of this method is to create a good instruction or prompt in which the models show good results in pre-training tests. To do this, we provide different examples to the models and ask the selected models what are the differences between critical and conspiratorial texts for subtask 1 and a definition of each element of oppositional narratives for subtask 2. We will feed the prompt with the information these models give us in their response, as we believe this information will help the model to detect each type of text or element. The used prompt are presented in Appendix A. To train the GPT-3.5 model, we use the OpenAI API with 1 epoch and to train LLaMA we use a method called QLoRA [10]. This approach facilitated a faster and more affordable process as it significantly reduced the hardware requirements. The model was loaded in 4 bits with the quantization data type NF4. As computational data type bf16 was used. Finally, LoRA update matrices were applied to the linear layers of the model. The LoRA rank was set to 16, the scaling factor (LoRA alpha) to 64 and the dropout to 0.05. We used a learning rate of 2e-4 and 1 example for the batch size, 10 epochs with an early stop of 3 epochs.

Furthermore, as we can see in Section 3.1 the dataset for subtask 2 is unbalanced, especially in the Spanish dataset where the class 'OBJECTIVE' appears fewer times. Since we do not have enough instances for the model to learn and considering this class inserts noise in the training of the models, we exclude this class during the training.

3. Experimental setup

3.1. Data

The dataset of this shared task [11] is composed of 10,000 messages of Telegram written in English and Spanish. These messages are related to the COVID-19 pandemic and labelled according to the annotation scheme for Task 1 and Task 2. The labels of Subtask 1 are CONSPIRACY and CRITICAL and the labels that we can associate to the span texts of the Subtask 2 are AGENT, FACILITATOR, VICTIM, CAMPAIGNER, OBJECTIVE, NEGATIVE_EFFECT. The dataset is divided into two splits, the first to train the developed systems and the second to test these systems.

The distribution of the train split of the dataset for Subtask 1: Distinguishing between critical and conspiracy texts, and Subtask 2: Detecting elements of the oppositional narratives are presented in Figures 1 and 2, respectively. We can observe that, for Subtask 1, the majority class is 'CRITICAL' although the dataset is not very unbalanced. Since Subtask 2 is a token-level classification, each text can have multiple labels and the same label can be repeated for each text instance. So, for Subtask 2, we can observe two figures, the first one (Subfigure 2a) represents the number of the texts in the dataset where the labels are. The label 'X' represents the texts where no label appears for the task. The second figure (Subfigure 2b) represents the number of times each label appears in the dataset. In each figure, we can see unbalanced data, for example, the minority class is 'OBJECTIVE' and for Spanish, this label appears in only 338 texts and fewer occurrences only 493 in front of 898 times of the same class appear in the English dataset with 1602 occurrences.

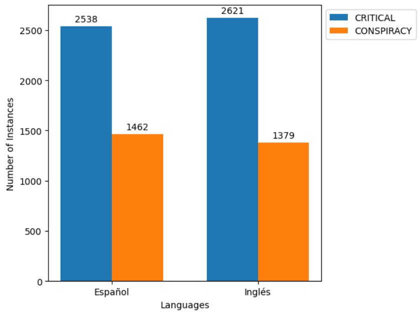
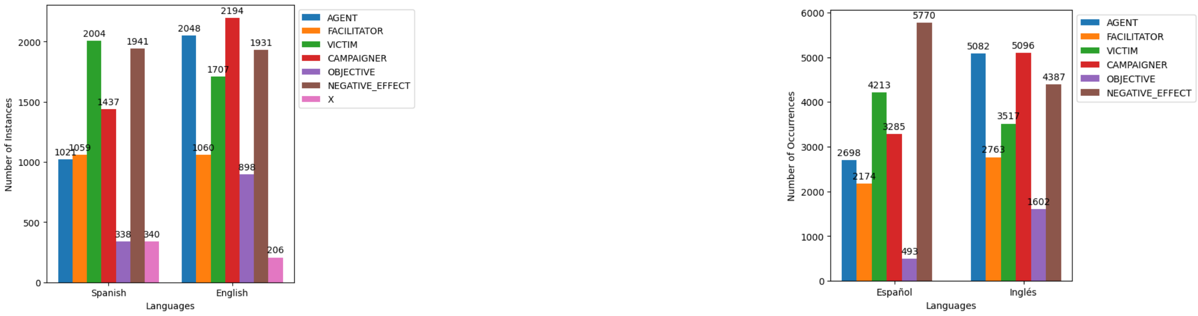


Figure 1: Distribution of the different classes presented in the Oppositional Thinking Analysis dataset for Subtask 1 (Distinguishing between critical and conspiracy texts).



(a) Number of instances of each class.

(b) Total of ocurrences of each class.

Figure 2: Distribution and number of occurrences of the different classes presented in the Oppositional Thinking Analysis dataset for Subtask 2 (Detecting elements of the oppositional narratives).

To carry out the experiments proposed in Section 3. We will divide the training set provided by the task organizers into three subsets. One to train the models, another to perform validation of our system during training, and finally a test one to evaluate how well our systems perform and select the best experiments to submit the final results to the task. The partitioning performed was done in a stratified way to maintain the same percentage of labels in the partitions created. The number of text instances of each class in the dataset for each subset can be seen in Table 1.

Table 1
Distribution of the number of instances of each class in the splits created to run our experiments.

Subtask	**Label**	**Spanish**			**English**		
		Train	**Validation**	**Test**	**Train**	**Validation**	**Test**
Subtask 1	CRITICAL	2284	127	127	2359	131	131
	CONSPIRACY	1316	73	73	1241	69	69
	Total	**3600**	**200**	**200**	**3600**	**200**	**200**
Subtask 2	AGENT	919	51	51	1843	102	103
	FACILITATOR	953	53	53	954	53	53
	VICTIM	1804	100	100	1536	86	85
	CAMPAIGNER	1293	72	72	1975	110	109
	OBJETIVE	304	17	17	808	45	45
	NEGATIVE_EFFECT	1747	97	97	1738	96	97
	Total	**3607**	**194**	**198**	**3594**	**201**	**205**

3.2. Experiments and Selected Models

To achieve the goal of the Oppositional Thinking Analysis shared task, we selected the following models: LLaMA3-8B-instruct [12], and GPT-3.5[13]. With these models, we want to study how the training of the models affects the achievement of the objectives of the proposed subtasks, and what differences exist between the size of the models, and their performance. Moreover, we propose two experiments for each task. Each experiment has a different configuration. The proposed experiments are the following:

- **Baseline.** This experiment employs the use of the model with a prompt strategy based on ZSL, providing reasoning for its responses. Our goal is to establish a reference experiment to evaluate the effectiveness of the proposed systems. In this case, we selected the GPT-3.5 model because we consider that a model with more parameters has more knowledge of the task and will be able to distinguish between the different classes of both subtasks without previous knowledge of them.
- **Fine-tuning.** This experiment applies techniques for efficient instruction learning of LLMs. To apply this experiment we are going to select the LLaMA3-8B-instruct model which is an open model and we have full control of its parameters and the GPT-3.5 model which belongs to a company and its use is not free. It also has more restricted parameters. To train LLaMA we will use

a technique for efficient learning of LLMs called QLoRA (Quantified Low-Rank Adaptation) [10]. This method accelerates the training process and makes it more accessible. QLoRA enables us to train models with a large number of hyperparameters using minimal hardware resources. This is achieved by not requiring the training of all model parameters and through the quantization of the numbers used during the training process. In this experiment, we load the selected model with 4 bits with the quantization data type NF4. The computational data type bf16 will be used. The LoRA update matrices were applied to the linear layers of the model. The LoRA rank is to be set to 16, the scale factor (LoRA alpha) to 64, and the dropout to 0.05. In addition, we used a learning rate of 2e-4, 1 example for the batch size, and 10 epochs with an early stop of 3 epochs. In the case of the GPT-3.5 model, we will use the openAI API and train the model with 1 epoch. For subtask 2, as seen in Section 3.1, we have unbalanced data, so we will propose two variants of this experiment:

- **FT_all**: Using all the labels in the dataset.
- **FT_withoutObjective**: Excluding the minority class. This class is 'OBJECTIVE'.

Because the use of GPT-3.5 is not free, to train the GPT-3.5 model for subtask 2 we only apply the fine tuning of the best variant of LLaMA training for subtask 2. As we can see in Section 4.1 the best variants for each language are to use all labels for English and to exclude the OBJECTIVE class for Spanish.

4. Results

In this section, we present the results obtained by the system developed as part of our participation in the "Oppositional thinking analysis" task. To evaluate our systems, we use the official metrics given by the organizers. Specifically, the MCC metric [14] that is a single-value classification metric which helps to summarize the confusion matrix or an error matrix. The MCC ranges between -1 and +1. A coefficient of +1 represents perfect prediction, 0 represents average random prediction and -1 represents inverse prediction. Moreover, for this task, the macro F1 score (harmonic mean of precision and recall for a more balanced summarization of model performance) and the specific macro F1 score for each class are provided. For subtask 2, the span F1 score [15] is used as the official metric. This metric calculates F1 measures per each class of the dataset and for each span identified. In addition, the organizers provide span recall and precision and the micro span F1 score. The experiments are conducted in two phases, the development phase, where we select the best models, and the evaluation phase where we evaluate the selected models and choose the best model to appear in the leaderboard of the evaluation campaign.

4.1. Development Phase

In order to select the best model for each subtask we trained the models selected in Section 3.2 with a subset of the train split provided by the organizers and evaluated them with other subsets of the train split. The results obtained in the development phase are shown in Tables 2, and 3.

Table 2 shows the results obtained in the experiments proposed above for subtask 1. As can be seen, the fine-tuning of LLaMA-8B-instruct shows promising results in all the metrics evaluated and obtains the best result when applied to English data. Its performance in Spanish is good, although it does not outperform the fine-tuning of GPT-3.5 model. This may be because LLaMA-8B-instruct does not have as extensive knowledge of Spanish as GPT-3.5, having been trained with more English data. In addition, the GPT-3.5 model shows greater consistency across languages by obtaining very similar results in both languages. If we look at the performance of the GPT-3.5 model with the ZSL experiment we see a big difference with the fitted models, as it does not achieve such promising results. This highlights the importance of adjusting the models to improve their performance in the proposed task.

The results of subtask 2 are shown in Table 3. For this subtask, we can see how LLaMA3-8B-instruct shows better results for English when trained with all classes compared to when trained without taking the minority class into account. However, for Spanish, the results of training with all classes show

Table 2
Results of the different experiments for Subtask 1 (distinguishing between critical and conspiracy texts) on test split of the train set of Oppositional Thinking Analysis. The selected model for the evaluation phase is shown in bold.

Model	Experiment	Lang.	MCC	F1-macro	F1-conspiracy	F1-critical
LLaMA3-8B-instruct	**FT_all**	**EN**	**0.7874**	**0.8930**	**0.8571**	**0.9288**
		ES	**0.6413**	**0.8204**	**0.7692**	**0.8716**
GPT-3.5	**FT_all**	EN	0.7345	0.8672	0.8261	0.9084
		ES	**0.7156**	**0.8552**	**0.8088**	**0.9015**
	ZSL	EN	0.4322	0.7014	0.6506	0.7521
		ES	0.4301	0.7143	0.6286	0.8000

lower performance than expected, being even below the ZSL strategy where the GPT-3.5 model has not been fitted to the task. Probably because having a very underrepresented class with few examples inserts noise during the model training process. For that reason, if we remove the minority class (OBJECTIVE) from Spanish we get a result more similar to what we would expect. As in the previous task, LLaMA3-8B-intruct performs better in English and GPT fits better in Spanish. If we look at the ZSL experiment we can see that this strategy is not effective for the task compared to model fitting, demonstrating the need to train the models to understand the differences between the different classes we have and where in the text they may appear.

Model	Experiment	Lang.	span-P	span-R	span-F1	micro-span-F1
LLaMA3-8B-instruct	**FT_withoutObjective**	EN	0.4810	0.4481	0.4461	0.4977
		ES	**0.4801**	**0.4068**	**0.4155**	**0.4862**
	FT_all	**EN**	**0.6013**	**0.4802**	**0.5140**	**0.5136**
		ES	0.2500	0.0024	0.0048	0.0050
GPT-3.5	**FT_all**	**EN**	**0.5225**	**0.4059**	**0.4532**	**0.4801**
	FT_withoutObjective	**ES**	**0.4806**	**0.3907**	**0.4282**	**0.5349**
	ZSL	EN	0.4246	0.1007	0.1563	0.1493
		ES	0.3740	0.7228	0.1147	0.1363

Table 3
Results of the different experiments for Subtask 2 (detecting elements of the oppositional narratives) on test split of the train set of Oppositional Thinking Analysis. The selected model for the evaluation phase is shown in bold.

4.2. Evaluation Phase

In the evaluation phase, we use the trained models of the development phase and evaluate them on the test set provided by the organizers. The systems submitted and their results for each run in subtasks 1 and 2 are presented in Tables 4, and 5 respectively.

Regarding subtask 1, because the LLaMA3-8B-instruct adjustment obtained the best results for English and because it was free, we decided to send these results in the 2 runs. In Spanish, we decided to send on one side the adjusted LLaMA3-8B-instruct model and on the other GPT-3.5. As can be seen in Table 4 we can see how the performance of the adjusted GPT-3.5 model outperforms the adjusted LLaMA3-8B-instruct model. This is not surprising since the same thing happened in the development phase and may be due to the fact that the prior knowledge that GPT has about Spanish is higher than that of LLaMA3-8B-instruct.

Table 4
Results of the different proposed strategies for Subtask 1 (distinguishing between critical and conspiracy texts) on Oppositional Thinking Analysis test set. The selected model for the leaderboard is shown in bold.

Run	Model	Experiment	Lang.	MCC	F1-macro	F1-conspiracy	F1-critical
Run 1	LLaMA3-8B-instruct	FT_all	EN	0.8297	0.9149	0.8886	0.9412
			ES	0.6780	0.8363	0.7841	0.8886
Run 2	**LLaMA3-8B-instruct**	**FT_all**	**EN**	**0.8297**	**0.9149**	**0.8886**	**0.9412**
	GPT-3.5		**ES**	**0.7429**	**0.8705**	**0.8319**	**0.9091**

The results of the systems submitted for subtask 2 can be seen in Table 5. For each submission we were allowed for this task we decided to submit an adjusted model with all classes for English and removing the minority class for Spanish. Since the differences between GPT-3.5 Spanish and LLaMA3-8B-instruct are minimal, we decided not to make combinations between these models for Spanish and to send the predictions made by each model for Spanish and English. As can be seen in this table, the model that best fits the task is LLaMA3-8B-instruct, probably because it has been trained with more epochs than GPT-3.5 and the task is somewhat more complex than the first one, since we have to choose between 6 classes and the parts in which it appears.

Table 5
Results of the different proposed strategies for Subtask 2 (detecting elements of the oppositional narratives) on Oppositional Thinking Analysis 2024 test set. The selected model for the leaderboard is shown in bold.

Run	Model	Experiment	Language	span-P	span-R	span-F1	micro-span-F1
Run 1	GPT-3.5	FT_all	English	0.5342	0.4243	0.4723	0.4945
		FT_withoutObjective	Spanish	0.4487	0.3674	0.4024	0.5149
Run 2	**LLaMA3-8B-instruct**	**FT_all**	**English**	**0.5553**	**0.4279**	**0.4582**	**0.4571**
		FT_withoutObjective	**Spanish**	**0.4630**	**0.4054**	**0.4151**	**0.4781**

Finally, we want to emphasize that the results obtained in both tasks by LLaMA3-8B-instruct are striking due to the large difference between the number of parameters that LLaMA3-8B-instruct has in comparison with GPT-3.5. This makes us think that as long as we have quality data and that they are representative of the classes, it is not so important to select models that are very large, since by training them a little more epochs we can obtain very similar and even better results to those with a large number of parameters.

4.3. Error Analysis

For each subtask, we present an error analysis of the final selected models in the test split used during our development phase.

For the first task, in Table 6 what we can see how difficult it is to recognize each of the labeled classes. For example in the first text for Spanish (id 9256) we can see how the comment is a criticism of the decision to change the brand at the time of putting the third dose of a vaccine, but also has part of conspiracy to say that to kill all carry the same thing, so the model assign the CONSPIRACY class. In the second example for Spanish (id 9076) we see a typical sentence of conspiracy theories ("they try to make us believe"), but the model is not able to detect it and thinks that it is more oriented to criticize how the different COVID variants are created. On the other hand, if we look at the English texts, we can see how just the conspiracy title of a thread of conversations where opinions are going to be exposed, already helps the model to classify it as CONSPIRACY instead of CRITICAL (id 151). Moreover, in the second English text (id 177), the purpose of the message is a conspiracy, but LLaMA3-8B-instruct model labels it as critical, probably because it thinks that is spreading an opinion of something that has been said in a podcast like AlexJonesShow.

On the other hand, in the texts related to Subtask 2, we find examples such as the ones shown in Table 7. If we look at the Spanish example (id 4263) as we have removed the OBJECTIVE class from the Spanish model, the model should not predict anything. However, it predicts various CAMPAIGNERS that are not even entities that promote something in the conspiracy. This suggests that the model is hesitant to recognize these types of entities. In the case of the English text, we can see how it is difficult for the model to recognize the negative effects that do not carry negations or negative words such as death. We can also see how it confuses the class CAMPAIGNER with FACILITATOR as in the case of "the " " scientific clerisy " " ".

Table 6
GPT-3.5 model for Spanish and LLaMA3-8B-instruct model for English error analysis for Subtask 1 (Distinguishing between critical and conspiracy texts). Examples of predictions from the test split created from the train split of Oppositional Thinking Analysis shared task dataset.

Model	Lang.	Id.	Text	Gold Label	predicted Label
GPT-3.5	ES	9256	AHORA TE DICEN QUE SI LA TERCERA DOSIS ES DE UNA MARCA DISTINTA A LAS PRIMERAS … ENTONCES ES MÁS EFICAZ (PARA MATAR QUERRÁN DECIR , TODAS LLEVAN LO MISMO) https :// www . infosalus . com / asistencia / noticia - administrar - tercera - dosis - vacuna - covid - 19 - compania - diferente - dos - primeras - eficaz - 20220425145746 . html (*NOW THEY TELL YOU THAT IF THE THIRD DOSE IS OF A DIFFERENT BRAND THAN THE FIRST … THEN IT IS MORE EFFECTIVE (TO KILL, THEY MEAN, THEY ALL CARRY THE SAME STUFF). https :// www . infosalus . com / asistencia / noticia - administrar - tercera - dosis - vacuna - covid - 19 - compania - diferente - dos - primeras - eficaz - 20220425145746 . html*)	CRITICAL	CONSPIRACY
		9076	Son los vacunados los que generan las variantes y los que contagian a los no vacunados , y no al revés como intentan hacernos creer (*It is the vaccinated who generate the variants and who infect the unvaccinated, and not the other way around as they try to make us believe.*)	CONSPIRACY	CRITICAL
LLaMA3 8B-instruct	EN	151	What Else Could They Have Lied to You About ? Tune into my conversation on Radical , with Maajid Nawaz … –> drtesslawrie . substack . com / p / on - what - else - could - they - have - lied I 'm delighted to share this wonderful conversation I had recently with Maajid Nawaz . Maajid is , amongst many things , a podcaster , an author with his own Substack here , and he was also a host at this year 's Better Way Conference . I really enjoyed speaking with him — he asks good questions — and we covered not just health but also the nefarious aims of the World Economic Forum and Big Pharma , the need for us to take control of our own health and also how to positively and practically prepare for challenging times ahead . Watch it here , and I hope you enjoy it . Have a wonderful Sunday , Tess Follow Me : –> @ audreywest	CRITICAL	CONSPIRACY
		177	# AlexJonesShow : It 's Official ! mRNA Covid Vaccines Are Euthanizing Thousands of Old People Worldwide ! - https :// ifw . io / hw8 Get Live Broadcast Alerts ! - Text : ' SHOW ' To : (833) 470 - 0222 $ 50 Off Alexapure Pro Water Filtration System : https :// www . infowarsstore . com / alexapure - pro - water - filtration - system	CONSPIRACY	CRITICAL

Table 7
LLaMA3-8B-instruct model error analysis for Subtask 2 (Detecting elements of the oppositional narratives). Examples of predictions from the test split created from the train split of Oppositional Thinking Analysis shared task dataset.

Lang.	Id.	Text	Gold Labels	predicted Labels
ES	4263	¿ LAS VACUNAS COVID INSTALARON « CARGAS ÚTILES » DE MARBURG QUE SERÁN LIBERADAS POR SEÑALES 5 G ? En la vacuna COVID - 19 se instaló nanotecnología que transportaba cargas útiles de virus quiméricos . No es ciencia ficción . El uso de señales externas para controlar implantes neurales usando nanotecnología está bien descrito en patentes y literatura médica . https :// ejercitoremanente . com / 2022 / 04 / 26 / las - vacunas - covid - instalaron - cargas - utiles - de - marburg - que - seran - liberadas - por - senales - 5 g / (*DID THE COVID VACCINES INSTALL MARBURG "PAYLOADS" TO BE RELEASED BY 5 G SIGNALS ? In the COVID - 19 vaccine was installed nanotechnology carrying chimeric virus payloads . It is not science fiction . The use of external signals to control neural implants using nanotechnology is well described in patents and medical literature . https :// ejercitoremanente . com / 2022 / 04 / 26 / las - vacunas - covid - instalaron - cargas - utiles - de - marburg - que - seran - liberadas - por - senales - 5 g /*)	{'text': 'El uso de señales externas para controlar implantes neurales usando nanotecnología', 'category': 'OBJECTIVE', 'chars': 225-307, *'english_text': ''* }	{'text': 'MARBURG','chars': 52-59, 'category': 'CAMPAIGNER'}, {'text': 'cargas útiles […] quiméricos','chars': 165-198, 'category': 'CAMPAIGNER', *'english_text': 'chimeric […] payloads',}*, {'text': 'El uso de […] literatura médica','chars': 225-358, 'category': 'CAMPAIGNER', *'english_text': 'The use […] medical literature'}*, {'text': 'ejercitoremanente','chars': 371-388, 'category': 'CAMPAIGNER'}, {'text': 'cargas - utiles - de - marburg','chars': 451-481, 'category': 'CAMPAIGNER'}
EN	11360	" Stanford professor who challenged lockdowns and ' scientific clerisy ' declares academic freedom ' dead ' - FOX NEWS After his life became a " " living hell " " for challenging coronavirus lockdown orders and the " " scientific clerisy " " during the pandemic , a medical professor at Stanford University claims that " " academic freedom is dead . " " SOURCE @ TheGreatResetTimes Follow us : Telegram | Chat Group | Twitter "	{'text': "Stanford professor who […] clerisy '", 'category': 'CAMPAIGNER', 'chars': 2-72}, {'text': "scientific clerisy '", 'category': 'FACILITATOR', 'chars': 52-72}, {'text': 'his life […] " " living hell " "', 'category': 'NEGATIVE_EFFECT', 'chars': 125-162}, {'text': 'the " " scientific clerisy " "', 'category': 'FACILITATOR', 'chars': 211-241}, {'text': 'a medical […] Stanford University', 'category': 'CAMPAIGNER', 'chars': 264-306}, {'text': 'TheGreatResetTimes', 'category': 'CAMPAIGNER', 'chars': 363-381}	{'text': 'a medical […] Stanford University','chars': 264-306, 'category': 'CAMPAIGNER'}, {'text': 'academic […] dead','chars': 323-347, 'category': 'NEGATIVE_EFFECT'}, {'text': 'TheGreatResetTimes','chars': 363-381, 'category': 'CAMPAIGNER'}, {'text': 'the " " scientific […] the pandemic','chars': 211-261, 'category': 'CAMPAIGNER'}

5. Conclusion

This paper presents the participation of SINAI research group in the Oppositional Thinking Analysis shared task at CLEF 2024. In the two subtasks, we explore how different fine-tuned LLMs (GPT-3.5 and LLaMA3-8B-instruct) perform using previous knowledge. For the first subtask, we have seen that GPT-3.5 model works better for Spanish than LLaMA3-8B-instruct model when fine-tuned to the task, while LLaMA3-8B-instruct performs better for English. In the second subtask, we found that LLaMA3-8B-instruct achieved better results than GPT-3.5 in both languages. We conclude that, in general, fine-tuning LLMs is effective for conducting oppositional thinking analysis tasks, especially when the number of classes is fewer. Furthermore, the good performance obtained by LLaMA3-8B-instruct demonstrates that it is not always necessary to use larger models; rather, we need models trained with quality data and given well-constructed input prompts so that they can effectively understand the task at hand. As future work, we plan to further analyze the misclassification of each class and provide the model with a complete definition to help in its reasoning. Additionally, since the detection of critical thinking is subjective, we aim to study how the classification of models is affected by texts with lower agreement among annotators and whether annotators' sociodemographic characteristics influence their reasoning. Finally, we want to investigate if the models are overfitted to the task data and if they perform well with other datasets.

Acknowledgments

This work has been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00), and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

References

[1] European Comission, Identifying conspiracy theories, https://commission.europa.eu/strategy-and-policy/coronavirus-response/fighting-disinformation/identifying-conspiracy-theories_en, Publication date unknown. Accessed: 13/06/2024.

[2] Oxford Learner's Dictionaries, Definition of critical thinking, https://www.oxfordlearnersdictionaries.com/definition/english/critical-thinking?q=critical+thinking, Publication date unknown. Accessed: 13/06/2024.

[3] A. A. Ayele, N. Babakov, J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, D. Moskovskiy, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, N. Rizwan, P. Rosso, F. Schneider, A. Smirnova, E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, X. Wang, M. Wiegmann, S. M. Yimam, E. Zangerle, Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification - Condensed Lab Overview, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association CLEF-2024, Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.

[4] M. Gambini, S. Tardelli, M. Tesconi, The anatomy of conspiracy theorists: Unveiling traits using a comprehensive twitter dataset, Comput. Commun. 217 (2024) 25–40. URL: https://doi.org/10.1016/j.comcom.2024.01.027. doi:10.1016/j.comcom.2024.01.027.

[5] A. Diab, R. Nefriana, Y.-R. Lin, Classifying conspiratorial narratives at scale: False alarms and erroneous connections, Proceedings of the International AAAI Conference on Web and Social Media 18 (2024) 340–353. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/31318. doi:10.1609/icwsm.v18i1.31318.

[6] M. Pustet, E. Steffen, H. Mihaljević, Detection of conspiracy theories beyond keyword bias in german-language telegram using large language models, 2024. arXiv:2404.17985.

[7] Y. Peskine, D. Korenčić, I. Grubisic, P. Papotti, R. Troncy, P. Rosso, Definitions matter: Guiding GPT for multi-label classification, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 4054–4063. URL: https://aclanthology.org/2023.findings-emnlp.267. doi:10.18653/v1/2023.findings-emnlp.267.

[8] Z. Liu, B. Liu, P. Thompson, K. Yang, S. Ananiadou, Conspemollm: Conspiracy theory detection using an emotion-based large language model, 2024. arXiv:2403.06765.

[9] D. Korenčić, B. Chulvi, X. B. Casals, M. Taulé, P. Rosso, F. Rangel, Overview of the oppositional thinking analysis pan task at clef 2024, in: G. Faggioli, N. Ferro, P. Galuvakova, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[10] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, 2023. arXiv:2305.14314.

[11] D. Korenčić, B. Chulvi, X. Bonet Casals, M. Taulé, P. Rosso, Pan24 oppositional thinking analysis, 2024. URL: https://doi.org/10.5281/zenodo.11199642. doi:10.5281/zenodo.11199642.

[12] AI@Meta, Llama 3 model card (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[13] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[14] D. Chicco, N. Tötsch, G. Jurman, The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation, BioData Mining 14 (2021). doi:10.1186/s13040-021-00244-z.

[15] G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, P. Nakov, Fine-grained analysis of propaganda in news article, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5636–5646. URL: https://aclanthology.org/D19-1565. doi:10.18653/v1/D19-1565.

A. Used Prompt

The prompts used for our experiment with ZSL and for tuning the selected models are presented in Table 8.

Subtask	Prompt
Subtask 1	You are an expert in the classification of critical and conspiratorial texts. Your task is to identify these CRITICAL and CONSPIRACY texts. CRITICAL messages criticize decisions made by an individual, a group of people, or a committee of experts. They may also expose personal concerns or opinions on an issue or decisions that have been made over time and are contradictory. Moreover, they make a claim about the theme, without delving into complex or implausible theories CONSPIRACY messages, on the other hand, see decisions as the result of a malevolent conspiracy by secret and influential groups. There are some differences between CRITICAL and CONSPIRACY messages: 1. Degree of Speculation: CRITICAL texts may contain unsubstantiated personal claims, but CONSPIRACY texts often go further by proposing complex and implausible theories. These theories lack solid evidence and are based on extreme speculation. 2. Level of Alarmism: CRITICAL texts may use alarming language. CONSPIRACY texts tend to be even more sensationalist and apocalyptic. They often include claims of impending catastrophic events or the existence of an ‘imminent danger’ that only the ‘awakened’ can see. 3. Global Conspiracy Tone: CRITICAL texts suggest specific concerns while CONSPIRACY texts often address much broader issues, such as the existence of a ‘secret world government’ or the manipulation of reality by unknown entities. Now you are going to receive a TEXT and based on everything explained above, argue your response, reasoning step by step, and put at the end of your answer the keyword ‘LABEL’ with the assigned class (CRITICAL or CONSPIRACY). TEXT: " "
Subtask 2	You are an expert in detecting elements of the texts. Since conspiracy narratives are a special type of causal explanation, your task consists in the recognition of text spans corresponding to the key elements of a text. Step 1: Identify all of the negative effects mentioned in the text and relate them to the oppositional narrative. A negative effect is a harmful consequence or negative impact related to conspiracy theories or critical aspects. Put these negative effects in the same form that they appear in the text in different lines with the keyword “NEGATIVE_EFFECT”. Step 2: Identify if there is an explicitly stated objective of the oppositional narrative. An explicit objective refers to a clear and direct statement outlining the goal or purpose of the narrative being presented. This objective is typically stated overtly within the text, providing insight into what the proponents of the narrative are trying to achieve or promote. Put these objectives in the same form that they appear in the text in different lines with the keyword “OBJECTIVE”. Step 3: Identify if there are victims of the oppositional texts. A victim is a specific individual or group that is negatively affected by the negative effects identified in step 1, harmful actions or policies described in the text. Put all victims with the keyword “VICTIM”. Step 4: Identify if there are conspirators in the text. A conspirator refers to the entity responsible for planning, executing, or supporting the main action or policy being discussed in the text. Moreover, a conspirator is responsible for the NEGATIVE_EFECTS Put all the conspirators identified with the keyword “AGENT”. Step 5: Identify if there is any facilitator in the text. A facilitator is a collaborator or entity that supports the agents in executing the main actions or policies discussed in the text. They assist in the achievement of the objectives outlined by the conspirators, often playing a role in enabling or promoting the negative effects on the victims. Put all the facilitators identified with the keyword “FACILITATOR”. Step 6: Detect the campaigners that appear in the text. A campaigner is an entity or someone who unmasks the conspiracy agenda, opposes the conspiracy narrative, and works to expose or challenge it. Moreover, a campaigner actively opposing the mainstream narrative and promoting his own opinion. Put all the campaigners identified with the keyword “CAMPAIGNERS”. Please answer each step with the exact part of the text and explain your answer for each step. If there is not a specific and clear element, do not provide it. TEXT: " "

Table 8

Used prompt for each subtask of the Oppositional Thinking Analysis shared task.