Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 127-138 recibido 03-12-2024 revisado 21-01-2025 aceptado 01-02-2025

FallacyES-Political: A Multiclass Dataset of Fallacies in Spanish Political Debates

FallacyES-Political: Un Dataset Multiclase de Falacias en Debates Políticos en Español

Fermín L. Cruz, Fernando Enríquez, F. Javier Ortega, José A. Troyano
Universidad de Sevilla, España
fcruz, fenros, javierortega, troyano@us.es

Abstract: Fallacies are pervasive in political discourse, shaping public opinion and influencing decision-making. Automatic detection and classification of fallacies is a challenging task, especially in non-English languages due to limited resources. In this study, we present FallacyES-Political, a novel dataset of fallacies extracted from 19 electoral debates held in Spain over three decades. The dataset comprises nearly 2,000 fallacies categorized into 16 types. To evaluate the dataset's utility, we conducted a comprehensive benchmarking of state-of-the-art Large Language Models (LLMs) in zero-shot classification. The results highlight the complexity of fallacy classification and the limitations of current LLMs in understanding context-dependent argumentation. Furthermore, we demonstrate the advantages of fine-tuning a compact, domain-specific model over relying on general-purpose LLMs, achieving notable improvements in classification accuracy with a more sustainable approach.

Keywords: Spanish Linguistic Resources, Political Discourse Analysis, Fallacy Classification.

Resumen: Las falacias son frecuentes en el discurso político, moldeando la opinión pública e influyendo en la toma de decisiones. La detección y clasificación automática de falacias es una tarea desafiante, especialmente en idiomas distintos del inglés debido a la escasez de recursos. En este estudio, presentamos FallacyES-Political, un novedoso conjunto de falacias extraídas de 19 debates electorales celebrados en España a lo largo de tres décadas. El conjunto de datos incluye casi 2.000 falacias categorizadas en 16 tipos. Para evaluar la utilidad del conjunto de datos, realizamos una evaluación comparativa de modelos de lenguaje de última generación (LLMs) en clasificación zero-shot. Los resultados destacan la complejidad de la clasificación de falacias y las limitaciones de los LLMs actuales para comprender argumentaciones dependientes del contexto. Además, demostramos las ventajas de ajustar un modelo compacto y específico para el dominio, en lugar de depender de LLMs de propósito general, logrando mejoras significativas en la precisión de la clasificación con un enfoque más sostenible.

Palabras clave: Recursos Lingüísticos en Español, Análisis del Discurso Político, Clasificación de Falacias.

1 Introduction

Logical, argumentative, or rhetorical fallacies (henceforth referred to as fallacies) are reasoning patterns that, despite their appearance of validity, are ultimately flawed (Tindale, 2007). The identification and classification of fallacies have been a philosophical and logical concern since antiquity, given their significant impact on public debates. They are often used to divert attention, reinforce stereotypes, or manipulate discourse, thereby generating confusion and misinformation. In the contemporary context, the spread of fake news, hate speech and political polarization have transformed the dynamics of public debate, making them a topic of active interest for researchers in the field of natural language processing (Rosso et al., 2020; Sepúlveda-Torres et al., 2024; Vallecillo-Rodríguez et al., 2024). Fallacy detection is particularly relevant in that context: in the digital ecosystem, these forms of faulty reasoning are instru-

ISSN 1135-5948 DOI 10.26342/2025-74-9

mentalized to amplify messages that, while lacking logical rigor, appeal to emotions and cognitive biases, thereby maximizing their impact. This phenomenon is especially evident in the rise of new populisms, whose rhetoric often relies on argumentative fallacies to oversimplify complex problems, construct common enemies, and consolidate polarizing narratives. Thus, the systematic study of fallacies is not only relevant from a logical perspective but also serves as a critical tool for understanding and countering manipulative strategies in contemporary politics.

In this study, we present FallacyES-Political, a dataset that compiles examples of fallacies extracted from electoral debates held in Spain over the past 30 years. We detail the methodology employed in the annotation process, which was designed to maximize quality and consensus in the annotations while addressing the intrinsic challenges and potential ambiguities associated with identifying fallacies. As an initial step to evaluate the practical utility of this resource, we assessed the performance of various general-purpose Large Language Models (LLM-based chatbots) in the automatic classification of fallacies, conducting a comparative analysis that highlights the inherent difficulties of this task and the challenges that remain to be addressed. As an alternative to relying on large general-purpose models, we also evaluated the fine-tuning of smaller models, demonstrating their practical utility for tackling this task with significantly lower energy and economic costs.

2 Background

There are several studies related to the use of fallacies in Spanish political discourse, highlighting how these rhetorical strategies are employed to shape public perception and consolidate ideological narratives. In the work of Sánchez García (2010), based on debates on the state of the nation held in Spain since the democratic transition, the logical structure and ethical implications of using fallacies are analyzed. The doctoral thesis by Fernández Barge (2021) delves into the impact of these strategies on legitimizing power and constructing political identities. Both works underscore the intersection between politics and the art of discourse, demonstrating how the use of fallacies not only manipulates public opinion but also reflects the sophistication or lack thereof in the argumentative design of speeches.

2.1 Resources

Regarding the development of resources on the use of fallacies in the political domain, it is worth noting the scarcity of studies addressing this task in Spanish. An interesting work is that of Benitez et al. (2022), which analyzes the speeches of candidates for the Mexican elections in 2006, 2012, and 2018. This work examines their argumentative structure by classifying propositions (conclusions and premises), though it focuses exclusively on one type of fallacy, Appeal to Emotion. In (Cruz et al., 2023), a resource is presented containing examples of fallacies in Spanish from two distinct domains or genres: approximately 2,000 examples of so-called *prototypical* fallacies, drawn from educational materials on teaching fallacies, and around 1,000 examples of *spontaneous* fallacies, which are examples extracted from user-written comments on a news website. The first set includes 12 different types of fallacies, while the second contains 8 types.

When considering languages other than Spanish, notable works include those by Habernal et al. (2017) and Habernal, Pauli, and Gurevych (2018), which explore the use of serious games as a tool for collecting fallacy examples. Other approaches rely on manual annotation of texts from various sources, such as discussion forums on Reddit (Habernal et al., 2018) or journalistic articles (Da San Martino et al., 2019). Sahai, Balalau, and Horincar (2021) also opt for the manual annotation of comments in Reddit forums, producing a corpus with over 3,000 examples classified into 8 types of fallacies. Jin et al. (2022) develop two distinct datasets: one based on prototypical examples of fallacies drawn from online educational resources and another derived from the manual annotation of news articles on the Climate Feedback website.

In the political domain, the work of Goffredo et al. (2022) presents a resource composed of approximately 1,600 fallacies of 6 types (and 14 subtypes) extracted from political debates in U.S. presidential elections up to 2016. In (Goffredo et al., 2023), around 200 new examples from the 2020 election debates are added.

2.2 Fallacy Detection and Classification

Habernal, Pauli, and Gurevych (2018) employed SVM algorithms and Bi-LSTM to tackle the task of fallacy classification, with the latter achieving the best results, obtaining an F1 score of 0.421 for six classes. Similarly, Habernal et al. (2018) utilized Bi-LSTM and convolutional neural networks (CNN) to classify texts, focusing specifically on binary detection of the 'ad hominem' fallacy, achieving a precision of 0.81. Da San Martino et al. (2019) adopted the BERT architecture (Devlin et al., 2018) with various configurations of final layers, addressing the classification task at different levels of granularity (document, paragraph, sentence, and word). This study classified 18 classes, although not all were fallacies, as it focused on analyzing propaganda techniques in news articles, achieving an F1 score of 0.6098 at the sentence level. Similarly, Sahai, Balalau, and Horincar (2021) applied approaches akin to those of Da San Martino et al. (2019), targeting the classification of eight types of fallacies in forum comments, with an F1 score of 0.5841. In the work by Jin et al. (2022), various transformer-based encoder and encoder-decoder models were evaluated, with Electra (Clark et al., 2020) standing out by achieving an F1 score of 0.5877.

In the political domain, Goffredo et al. (2022) achieved an F1 score of 0.74 in the task of token-level fallacy labeling across six fallacy categories. While this represents a promising result, it is important to note that the model benefited from the availability of argumentative features, including claims, premises, and their relations, which provided additional structural information crucial for improving detection accuracy.

Lastly, Cruz et al. (2023) is the only work reporting fallacy classification results in Spanish, achieving an F1 score of 0.6775 for so-called prototypical fallacies (12 classes) and 0.6385 for spontaneous fallacies (8 classes). In both cases, these results were obtained through fine-tuning RoBERTa-base-BNE model (Gutiérrez Fandiño et al., 2022).

3 *Dataset*

The FallacyES-Political dataset consists of fallacy examples extracted from 19 debates between candidates for Spain's General Elections, held on nationally broadcast radio or television (see Table 1). All such debates up to the 2023 General Elections were processed[1]. The debates were transcribed using WhisperX (Bain et al., 2023) and processed by three annotators with backgrounds in journalism and philosophy (see Section 3.1 for further details).

The resulting dataset comprises 1,965 instances drawn from the speeches of 33 representatives of 11 different political parties. For each instance, the dataset includes the text, the type of fallacy, the debate from which it was extracted, and the speaker's identity. Additionally, the context surrounding the excerpt is provided, as it is sometimes essential for correctly determining the fallacy type. The resource is publicly available at *https://zenodo.org/records/14836328*.

Table 3 illustrates examples for the 16 types of fallacies included in the dataset, defined as follows:

Ad Hominem (AH): Insults or attacks the opponent instead of confronting and developing the argument aimed at defending or rejecting a proposal, or discredits the opponent's proposal by referring to past circumstances or facts that would disqualify and discredit them from acting.

Ad Populum (AP): Bases the truth (or falsity) of an argument on the fact that most people believe it to be true (or false). Sometimes, the proponent speaks on behalf of a group and generalizes to present their opinion or proposal as common sense, unquestionable.

Appeal to Authority (AA): Mentions the name of an alleged authority (person, organization, or group) who agrees with the claim or whose actions support it, without providing concrete evidence beyond merely citing the authority.

Appeal to Emotion (AE): Adds unnecessary or exaggerated emotional language to the argument to exploit the audience's emotional response—such as pity, anger, love—to prevent rational thinking and prompt uncritical acceptance of the claim.

Appeal to Fear (AF): A subtype of appeal to emotions, in this case to fear. It seeks to convince the audience that if they do not accept the claim or act in a certain way, a

[1] With the exception of two debates for which access to recordings was unavailable: one from 1993 and another from 2015. Note that debates were not held during every election cycle.

Elect.	Date	Participants	Organizer
1993	24 May	J.M. Aznar (PP), F. González (PSOE)	Antena 3
2008	25 Feb.	M. Rajoy (PP), J.L.R. Zapatero (PSOE)	AcademiaTV
	3 March	M. Rajoy (PP), J.L.R. Zapatero (PSOE)	AcademiaTV
2011	7 Nov.	M. Rajoy (PP), A.P. Rubalcaba (PSOE)	AcademiaTV
2015	23 Nov.	P. Iglesias (Podemos), A. Rivera (Cs)	Univ.Carlos III
	30 Nov.	P. Sánchez (PSOE), P. Iglesias (Podemos), A. Rivera (Cs)	El País
	14 Dec.	P. Sánchez (PSOE), M. Rajoy (PP)	Atresmedia-AcademiaTV
2016	13 June	M. Rajoy (PP), P. Sánchez (PSOE), P. Iglesias (UP), A. Rivera (Cs)	AcademiaTV
2019 (April)	22 April	P. Sánchez (PSOE), P. Casado (PP), P. Iglesias (UP), A. Rivera (Cs)	RTVE
	23 April	P. Sánchez (PSOE), P. Casado (PP), P. Iglesias (UP), A. Rivera (Cs)	Atresmedia
	16 April	C. Álvarez de Toledo (PP), M.J. Montero (PSOE), I. Montero (UP), I. Arrimadas (Cs), G. Rufián (ERC-Sobiranistes), A. Esteban (PNV)	RTVE
	20 April	T.García Egea (PP), F. Sicilia (PSOE), A. Garzón (UP), T. Cantó (Cs), G. Rufián (ERC-Sobiranistes), L. Borràs (JxCAT), A. Esteban (PNV)	LaSexta
2019 (Nov.)	4 Nov.	P. Sánchez (PSOE), P. Casado (PP), P. Iglesias (UP), A. Rivera (Cs), S. Abascal (Vox)	AcademiaTV
	1 Nov.	A. Lastra (PSOE), C. Álvarez de Toledo (PP), I. Arrimadas (Cs), I. Montero (UP), I. Espinosa de los Monteros (Vox), G. Rufián (ERC), A. Esteban (PNV)	RTVE
	2 Nov.	F. Sicilia (PSOE), C. Gamarra (PP), M. Rodríguez (Cs), N. Vera (UP), J.O. Smith (Vox), G. Rufián (ERC), A. Esteban (PNV), L. Borràs (JxCAT)	LaSexta
	7 Nov.	M.J. Montero (PSOE), A. Pastor (PP), I. Arrimadas (Cs), I. Montero (UP), R. Monasterio (Vox)	LaSexta
2023	10 July	P. Sánchez (PSOE), A.N. Feijóo (PP)	Atresmedia
	13 July	P. López (PSOE), C. Gamarra (PP), A. Vidal (Sumar), I. Espinosa de los Monteros (Vox), A. Esteban (PNV), O. Matute (EH Bildu), G. Rufián (ERC)	RTVE
	19 July	P. Sánchez (PSOE), Y. Díaz (Sumar), S. Abascal (Vox)	RTVE

Table 1: Debates for Spanish General Elections in FallacyES-Political.

catastrophe or disaster will occur.

Complex Question (CQ): Poses a question that contains an implicit assertion or unproven premise, so that answering the question implies acceptance of the implicit assertion.

False Analogy (FA): Compares elements or situations that are not comparable (or at least fails to provide enough arguments to make them so), projecting the characteristics of one element onto the other and drawing conclusions based on this.

False Cause (FC): Confuses correlation with causation by explaining a complex event based on a single or few factors, ignoring other factors that could influence the outcome.

False Dilemma (FD): Presents two or a few options as the only possible ones (or explicitly presents only one option, implying a single opposing option implicitly), when there are actually many possible options.

Flag Waving (FW): A subtype of appeal to emotions that appeals to a sense of belonging to a group, so that an audience identifying with that group accepts the arguments without question.

Hasty Generalization (HG): Draws a general conclusion based on one or few cases, moving from the anecdotal to the categorical.

Poisoning the Well (PW): A strong type of Ad Hominem. It involves a lengthy sequence of negative accusations, or a single but extremely harsh accusation, aimed at an adversary or group to discredit or ridicule

Type	Instances	Type	Instances
SM	351	CQ	74
FC	304	FA	67
AH	285	HG	59
FD	181	AP	50
AE	155	RH	28
PW	152	SS	23
AF	115	FW	19
AA	87	SG	15
Total Instances			1965

Table 2: Number of instances for each type of fallacy in the dataset.

everything they subsequently say.

Red Herring (RH): Introduces a new topic unrelated to the original debate, distracting attention from the original issue.

Slippery Slope (SS): Suggests an improbable, exaggerated outcome that might occur as a result of a particular action. Intermediate premises are usually omitted, using an initial premise as a first step toward an exaggerated claim.

Slogan (SG): A brief and impactful phrase used to excite the audience, often accompanied by another fallacy called the argument by repetition.

Strawman (SM): Reformulates the opponent's arguments or past actions in an exaggerated, simplified, or caricatured way, then proceeds to attack this new distorted version of the arguments.

Table 2 presents the number of instances for each type of fallacy, while Figure 1 depicts the distribution of text and context lengths for these instances. Overall, there are no significant differences in average text lengths across fallacy types, except for instances of Poisoning the Well, which tend to be considerably longer due to their nature.

3.1 Annotation Methodology

The use of a fallacy does not imply the falsity of a conclusion but rather a flawed argumentative structure. In fallacy annotation, the focus is not on the truthfulness of conclusions but on the robustness of the reasoning. It is crucial to distinguish between fallacious arguments and valid opinions, even when these are not scientifically irrefutable. Given that the boundary between fallacious and non-fallacious reasoning can be subjective, annotation is a complex task. To address this challenge, strategies such as annotator training, clear criteria, concrete examples, and iterative rounds of review were implemented, along with fostering consensus through discussions in cases of disagreement. This ensured a consistent and well-founded analysis. The annotations were carried out by three annotators with backgrounds in journalism and philosophy and were supervised and validated by the authors of this study.

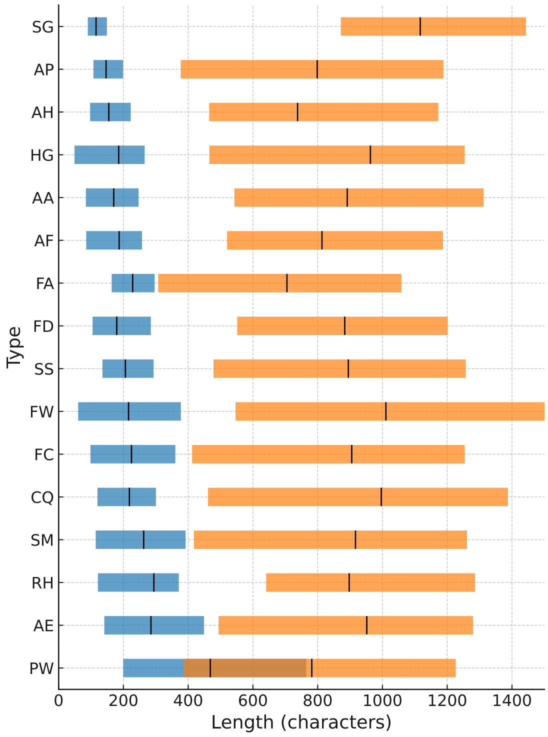

Figure 1: Box plot comparing the distribution of text (blue) and context (yellow) lengths across different types of fallacies. The boxes represent the Q1-Q3 range, with medians marked as black lines.

The first step was to create a catalog of potential fallacy types, including definitions and examples, based on those proposed in the literature (Cruz et al., 2023; Goffredo et al., 2022; Jin et al., 2022). Using this catalog, the three annotators processed the transcripts of two complete debates, attempting to identify as many fallacies as possible from the cataloged types. In subsequent meetings, the identified fallacies were analyzed and discussed, and the 16 types ultimately included in the dataset were selected. An initial version of the annotation guide was developed,

Type	**Example**
Ad Hominem	*(...) Marxism is to blame for 150 million deaths and* the Marxist apprentices who govern us today are to blame for many things (...)
Ad Populum	(...) this is what is done in all countries of the world, this is normal, sensible, and reasonable (...)
Appeal to Authority	I would create a tax on banks, as has been created in Great Britain, or Germany, or France, or Sweden (...)
Appeal to Emotion	That's why I ask you to vote with confidence. I ask you to vote with hope and to vote for the Socialist Party.
Appeal to Fear	3,300,000 unemployed people feel anxiety and worry because they do not know if they will be able to find a job again. Others who are employed also feel anxiety and worry because they don't know if they will be among the 2,950 Spaniards who lose their job every day.
Complex Question	*(...) he began by saying that the concept of nation is debatable and questionable (...) he opened a Pandora's box that no one was asking for. He has created conflicts between autonomous communities over heritage, funding, and investments (...) he has divided Spaniards (...)* Could you explain why you did all this and if you truly believe that Spain is now more united and cohesive than ever?
False Analogy	Everyone says they are going to fix things like magic. Tsipras in Greece said the same, and all that ended with pensions falling, the corralito, etc.
False Cause	The labor reform of the Popular Party has caused there to be poor workers who cannot make ends meet.
False Dilemma	We are going to have to choose between democracy or repression, between having legitimate representatives in prison or in exile or occupying their seats.
Flag Waving	I believe that Spain and Spaniards have a great future ahead of us. We have a good foundation. We are a great nation.
Hasty Generalization	Since then, they have not supported any of the laws that have extended rights in Spain. Not the gay marriage law, nor the abortion law, nor the one in 1985, nor the current one. They have been against the equality law, against all the laws.
Poisoning the Well	Sánchez's obsession, do you know what it has been? How to go down in history. And what Sánchez is going to do will go down in history as the prime minister who passed a law that benefited rapists, sexual aggressors, pedophiles, and even murderers, the "only yes is yes" law.
Red Herring	*[MODERATOR] Where does Spain stand in this new Europe and its relationship with Latin America? And in which direction will you lead it if you become president of the Government? [CHANGE TO SPEAKER_09]* Foreign policy is already national policy. European policy is national policy. What happens, for example, or what happened in Kabul, and of course all my recognition to the national police officers and all my solidarity with the families of the two officers killed in Kabul (...)
Slippery Slope	If we don't put an end to this immediately, what we are seeing in Catalonia today we will see in Navarra, in the Valencian Community, in the Balearic Islands, and in Catalonia in 10, 15 years. And that threatens the equality of all Spaniards and the survival of Spain.
Slogan	When we are together, we are unstoppable. When we are together, we are stronger.
Strawman	You, on the other hand, only consider the possibility of, in short, suffocating families and companies with more and more taxes, which are then squandered and unfortunately wasted.

Table 3: Examples of fallacies extracted from FallacyES-Political (originally in Spanish). Some examples include part of the context in italics.

which was refined during periodic meetings throughout the annotation process.

The remaining 17 debates were annotated independently by two annotators for each debate, aiming to maximize the coverage of identified fallacies. After the annotations were completed, a cross-validation process was conducted to assess their reliability. In this process, each annotator was presented with the text segments annotated by the others, along with their contexts, and was asked to select the appropriate type of fallacy. The average agreement rate was 47.67%. Analysis of the discrepancies during meetings with both annotators and researchers revealed several causes.

First, the definitions of some fallacy types occasionally overlapped (especially for types that are subcategories of others, such as Poisoning the Well and Appeal to Fear). Efforts were made to refine these definitions in the annotation guide. Second, identifying certain fallacy types requires a degree of subjective judgment from the annotator. For example, determining whether an argument involves false causality may depend on the annotator's judgment of whether the supposed causes adequately explain the consequence or whether the event is too complex to be explained so simply. Third, some text segments could contain more than one fallacy. To address this, segments were narrowed as much as possible to minimize overlaps, and a multi-label annotation approach was adopted. This allowed annotators to assign more than one fallacy type to a text segment when non-overlapping segments could not be identified[2].

With these updated instructions, meetings were held to resolve disagreements from the earlier stage. Finally, the researchers performed a final curation process by excluding any instances that raised doubts about the instructions in the annotation guide, and making the text segments as concise as possible.

4 Experimentation

To evaluate the practical utility of the dataset, we designed two sets of experiments that explore different approaches to the automatic classification of fallacies. First, we analyzed the performance of various general-purpose Large Language Model (LLM) chatbots, conducting a comparative analysis that highlights the inherent difficulties of this task and the challenges that remain. Second, we explored a more cost-effective and sustainable alternative by fine-tuning smaller models, demonstrating their capability to effectively tackle the task with significantly reduced energy and economic costs.

4.1 Benchmarking LLMs as Zero-Shot Fallacy Classifiers

Contemporary Large Language Models have demonstrated remarkable capabilities in addressing natural language processing tasks that were previously considered intractable. With each new generation of models, performance on standard benchmarks continues to improve significantly. However, increasing reliance on these benchmarks introduces a critical issue: their reliability diminishes over time. As models achieve higher scores on these evaluations, their results often reflect superficial correlations rather than deep understanding or advanced reasoning abilities. This stems, in part, from the use of benchmarks to guide model design and optimization, which can lead to overestimation of their true capabilities and inadequate differentiation among models.

The classification of argumentative fallacies is a particularly complex task, as evidenced by the labor-intensive annotation process undertaken for the creation of our dataset. Accurate identification of fallacies requires a profound understanding of the context in which they appear, sufficient world knowledge, and familiarity with the linguistic and rhetorical nuances that can influence their interpretation. This complexity makes fallacy classification an ideal challenge for assessing LLM performance in zero-shot scenarios, where models lack prior task-specific training.

In this section, we present a series of experiments designed to evaluate the performance of several leading LLMs in the classification of argumentative fallacies. To this end, we randomly selected 50 instances for each of 12 fallacy types (excluding four types with fewer than 50 instances for these experiments, see Table 2). We generated a system prompt that defined the task, including the definitions of the fallacy types to be classified (see Figure 2). For each instance, we created a user prompt containing the fallacy text and, in some experiments, the context in which it appeared. All experiments were conducted with a temperature setting of 0, maximizing reproducibility, and a maximum output length of 10 tokens.

Although the instructions provided to the models explicitly requested that they respond only with the name of one of the defined fallacy types, some outputs did not fully adhere to this format. In such cases, we assigned the predicted class as the type of fallacy with the smallest edit distance to the output among the defined types. Table 4 presents the F1-scores obtained using this approach, both globally and for each fallacy type, across dif-

[2] In the final version of the dataset, only 11 instances have two labels. These instances were excluded from the experiments described in Section 4.

ferent language models selected from the top-ranked models in the Chatbot Arena LLM Leaderboard[3] (Chiang et al., 2024), a platform that evaluates and ranks LLMs through anonymous comparisons and user voting, using the Elo rating system.

4.1.1 Results

For each model, we report results both with and without the inclusion of context in the user prompts to assess its importance for accurate fallacy classification. Except for the llama-70b model, which was run locally on our infrastructure, all other models were accessed via paid APIs provided by their respective organizations.

```
The user will provide a TEXT extracted
    from an electoral debate. The TEXT
    will be accompanied by its CONTEXT,
    both prior and subsequent. The TEXT
    contains some type of fallacy. The
    system must determine the type of
    fallacy from the following categories
    :
<CATEGORIES>
* Ad Hominem: Insults or attacks the
    opponent instead of confronting...
(rest of definitions omitted)
</CATEGORIES>

The system MUST ONLY look for fallacies
    in the TEXT, NOT in the CONTEXT. The
    system MUST ONLY respond with one of
    the categories listed above, without
    providing any additional explanation.
     The system must understand that the
    TEXT may match more than one category
    , but it must select the category
    that best fits the TEXT according to
    the provided definitions. In both the
     TEXT and the CONTEXT, changes in
    speaker turn are indicated with the
    tag [CHANGE TO SPEAKER_UID], where
    SPEAKER_UID is the identifier of the
    speaker taking the floor.
```

Figure 2: System prompt used in zero-shot classification experiments (originally in Spanish).

The results confirm the complexity of the task of fallacy classification. Nevertheless, it is noteworthy that all models significantly outperform the expected F1-score for random predictions (0.08 for a 12-class classification task). Some models, such as GPT-4o, Claude-3.5, and Gemini-1.5, achieve a weighted F1-score exceeding 0.5. These results suggest that the models are capable of interpreting the provided definitions and identifying relevant patterns in the texts, even without prior training on the dataset.

GPT-4o emerges as the top-performing model in this study, achieving the highest overall score and excelling in 6 of the 12 fallacy categories. The performance gap between GPT-4o and other models indicates that GPT-4o may possess greater capacity to handle complex instructions. The inclusion of additional context proves beneficial for the best-performing models (GPT-4o and Claude-3.5), whereas other models, such as Llama-70b and Nemotron, exhibit a pronounced decline in performance when context is incorporated. Although detailed parameter information for all models is not publicly available, it seems that smaller models face greater challenges in integrating contextual information and adhering strictly to prompt instructions. Specifically, these models may overinterpret the context, misapplying it to the classification of the target text.

The most challenging fallacies to classify are, in descending order, Strawman, False Analogy, and Hasty Generalization. This difficulty could be attributed to the more abstract and less defined nature of these categories, where the boundary between what constitutes a fallacy and what does not is more subjective or dependent on broader semantic context. Conversely, the best-recognized categories, such as Complex Question, Appeal to Authority, and Ad Populum, exhibit more evident and straightforward argumentative patterns that the models can identify with greater ease.

4.1.2 Misclassification Analysis

Figure 3 illustrates the most frequent confusions for each fallacy type, based on the outputs of the best-performing experiment. Although it is not possible to explain all errors in this way, many confusions arise from shared rhetorical or structural components: Ad Hominem, Poisoning the Well, and Strawman are centered on attacking the opponent, although they differ in intensity and focus, with Poisoning the Well being more aggressive and Strawman characterized by a caricatured distortion of arguments. Appeals to Authority include comparisons with countries considered as authorities (see Appels to Authority example in Table 3), which makes it logical for the classifier to confuse

[3] Some models were excluded due to lack of API availability or costs that made the experiments infeasible

Model	Ctx	Global	AH	AP	AA	AE	AF	CQ	FA	FC	FD	HG	PW	SM
GPT-4o	Yes	**0.570**	0.50	0.65	0.64	**0.70**	**0.65**	**0.82**	**0.46**	0.51	**0.57**	**0.48**	0.49	0.37
	No	0.559	0.48	**0.74**	0.70	0.60	0.63	0.81	0.45	0.46	0.55	0.44	**0.50**	0.34
Claude-3.5	Yes	0.537	**0.53**	0.62	0.72	0.65	0.57	0.72	0.44	**0.53**	0.46	0.43	0.42	0.36
	No	0.528	0.52	0.63	0.69	0.59	0.57	0.69	0.43	0.49	0.52	0.40	0.45	0.37
Gemini-1.5	Yes	0.496	0.43	0.61	0.75	0.59	0.51	0.72	0.41	0.51	0.44	0.40	0.21	0.36
	No	0.513	0.42	0.69	0.77	0.59	0.49	0.78	0.30	0.51	0.53	0.38	0.34	0.36
Llama-405b	Yes	0.436	0.39	0.52	0.74	0.55	0.38	0.51	0.35	0.15	0.50	0.43	0.43	0.28
	No	0.422	0.34	0.49	**0.78**	0.53	0.37	0.59	0.33	0.14	0.41	0.34	0.41	0.32
Grok-b	Yes	0.408	0.38	0.55	0.72	0.60	0.51	0.18	0.45	0.38	0.38	0.43	0.04	0.27
	No	0.431	0.37	0.62	0.67	0.63	0.56	0.53	0.38	0.41	0.37	0.36	0.11	0.18
Nemotron	Yes	0.311	0.31	0.49	0.70	0.45	0.46	0.15	0.18	0.35	0.30	0.07	0.00	0.28
	No	0.420	0.36	0.60	0.72	0.47	0.31	0.60	0.30	0.41	0.44	0.35	0.11	0.36
Llama-70b	Yes	0.318	0.31	0.47	0.64	0.42	0.46	0.15	0.20	0.31	0.27	0.33	0.00	0.25
	No	0.405	0.35	0.60	0.74	0.52	0.28	0.56	0.17	0.39	0.38	0.40	0.08	**0.38**

Table 4: F1-weighted results for the zero-shot classification task using some of the best instruction-tuned language models available, evaluated on a dataset of 50 instances per fallacy type. The models used include gpt-4o-2024-08-06 (GPT-4o), claude-3-5-sonnet-20241022 (Claude-3.5), gemini-1.5-pro-001 (Gemini-1.5), meta/llama-3.1-405b-instruct (Llama-3.1-405b), x-ai/grok-beta (Grok-b), nvidia/llama-3.1-nemotron-70b-instruct (Nemotron), and lmstudio-community/Meta-Llama-3.1-70B-Instruct-GGUF (Llama-70b). The results are presented in descending order of overall performance.

them with False Analogy. Hasty Generalization, False Cause, Ad Populum, and False Dilemma are often based on simplistic reasoning or weak connections between premises and conclusions, making them harder to distinguish. Finally, Appeals to Emotion and Appeals to Fear share an emotional focus, with the latter being a more specific subtype of the former.

4.2 Fine-Tuning Experiments

We reproduced the experimental setup described in Cruz et al. (2023), performing fine-tuning on RoBERTa-base-BNE (Gutiérrez Fandiño et al., 2022), a 125M-parameter encoder-only language model pretrained on the BNE corpus, a 540GB collection of Spanish texts. Being an encoder-only model, a dense output layer with as many neurons as output classes (12) was added for the fine-tuning process.

Two runs were conducted, with and without context, using 90% of the available instances for training and the remaining 10% for evaluation, randomly distributed in a stratified manner. The same hyperparameter values as in Cruz et al. (2023) were used (e.g., a learning rate of 5e-5 and a batch size of 16).

4.2.1 Results

Table 5 presents the F1-scores obtained from these experiments. For comparative purposes, the table also includes the results reported in Cruz et al. (2023) for prototypical and spontaneous fallacies, as well as the F1-score achieved by the zero-shot classification approach using GPT-4o on the same 10% evaluation set employed in the fine-tuning experiments.

Model	Dataset	#class	#train	Ctx	F1
BNE	prototypical	12	1874	No	0.678
	spontaneous	8	830	No	0.639
	political	12	1858	No	0.641
				Yes	0.519
GPT-4o	political	12	-	No	0.554
				Yes	0.575

Table 5: F1-weighted results for the supervised classification task using the RoBERTa-base-BNE (BNE) model fine-tuned on the different sections of the FallacyES resource. The results of the zero-shot classification task using the GPT-4o model on FallacyES-Political are also presented.

The best result in supervised fallacy classification within the political domain (F1 = 0.641) was achieved without using context. When context was included, the classi-

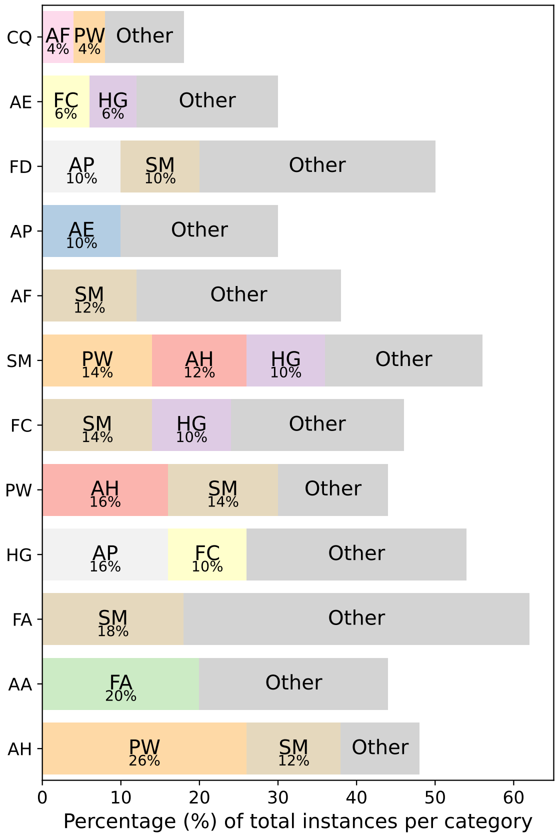

Figure 3: Confusions by fallacy type for the zero-shot classifier based on GPT-4o with context. For each type, only the most common confusions are shown.

fier's performance declined, which could be attributed to the model's relatively small size and the limited number of training instances available. The performance observed in the political domain is comparable to that achieved for spontaneous fallacies (extracted from online discussions among users on a news website). However, the new experiments used a larger number of training instances, suggesting that the political domain may present greater difficulty. Nonetheless, the differences are small and do not allow for definitive conclusions.

Despite its inability to effectively leverage context, the performance achieved through fine-tuning RoBERTa-BNE is noticeably better than that of GPT-4o in the zero-shot classification approach. This result was obtained using a compact, open-source model that can be run locally on a consumer-grade GPU. We believe this demonstrates the value of employing smaller, task-specific models for challenges like fallacy classification, rather than relying on large, general-purpose models. While large models are versatile, they entail significantly higher computational, energy, and economic costs, as well as dependence on third-party infrastructures. This underscores the practicality and sustainability of fine-tuning smaller models tailored to specific tasks.

5 Conclusions

In this study, we presented FallacyES-Political, a dataset comprising fallacies extracted from Spanish political debates. The labor-intensive annotation process and the experimental results highlight the inherent difficulty of fallacy classification. Despite the advanced capabilities of some general-purpose LLMs, fine-tuning a compact model proved to be more effective. These findings underscore the importance of addressing complex linguistic challenges with methodologies that combine task-specific resources and fine-tuned models, offering a more sustainable alternative to the use of large general-purpose models.

Given the promising results obtained with some of the models tested, we plan to explore the semi-supervised expansion of the dataset by manually selecting fallacy candidates from new examples that receive consistent classifications across multiple models. Additionally, we are interested in the automatic generation of narratives explaining why a specific argument is classified as a fallacy. This approach would not only enhance interpretability but could also serve as an educational tool to promote critical thinking among audiences. Furthermore, such explanations could be integrated into automated moderation systems for user-generated content on social media platforms.

Acknowledgments

This publication is part of the project PID2021-123005 funded by MCIN/AEI /10.13039/501100011033/ and by FEDER, UE.

References

Bain, M., J. Huh, T. Han, and A. Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Benitez, K. N., N. A. Castro-Sánchez, H. J. Salazar, and G. Bel-Enguix. 2022. Corpus de falacias por apelación a las emociones: una aproximación a la identificación automática de falacias. *Linguamática*, 14(2):59–72.

Chiang, W.-L., L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Clark, K., M.-T. Luong, Q. V. Le, and C. D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Cruz, F. L., J. A. Troyano, F. Enríquez, and F. J. Ortega. 2023. Detección y clasificación de falacias prototípicas y espontáneas en español. *Procesamiento del Lenguaje Natural*, 71:53–62.

Da San Martino, G., S. Yu, A. Barrón-Cedeno, R. Petrov, and P. Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5636–5646.

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

Fernández Barge, X. 2021. *Un análisis crítico del uso de falacias y mecanismos de refuerzo y atenuación en el discurso de los líderes políticos españoles en debates televisivos*. Ph.D. thesis, Universidad de Cádiz.

Goffredo, P., M. Chaves, S. Villata, and E. Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112, Singapore, December. Association for Computational Linguistics.

Goffredo, P., S. Haddadan, V. Vorakitphan, E. Cabrio, and S. Villata. 2022. Fallacious argument classification in political debates. In *IJCAI*, pages 4143–4149.

Gutiérrez Fandiño, A., J. Armengol Estapé, M. Pàmies, J. Llop Palao, J. Silveira Ocampo, C. Pio Carrino, C. Armentano Oller, C. Rodriguez Penagos, A. Gonzalez Agirre, and M. Villegas. 2022. MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, 68.

Habernal, I., R. Hannemann, C. Pollak, C. Klamm, P. Pauli, and I. Gurevych. 2017. Argotario: Computational argumentation meets serious games. *arXiv preprint arXiv:1707.06002*.

Habernal, I., P. Pauli, and I. Gurevych. 2018. Adapting serious game for fallacious argumentation to german: Pitfalls, insights, and best practices. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Habernal, I., H. Wachsmuth, I. Gurevych, and B. Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. *arXiv preprint arXiv:1802.06613*.

Jin, Z., A. Lalwani, T. Vaidhya, X. Shen, Y. Ding, Z. Lyu, M. Sachan, R. Mihalcea, and B. Schölkopf. 2022. Logical fallacy detection. *arXiv preprint arXiv:2202.13758*.

Rosso, P., F. Casacuberta, J. Gonzalo, L. Plaza, J. Carrillo, E. Amigó, M. F. Verdejo, M. Taulé, M. Salamó, and M. A. Martí. 2020. Mismis: Misinformation and miscommunication in social media: aggregating information and analysing language. *Procesamiento del lenguaje natural*, 65:101–104.

Sahai, S. Y., O. Balalau, and R. Horincar. 2021. Breaking down the invisible wall of informal fallacies in online discussions. In *ACL-IJCNLP 2021-Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Sánchez García, F. J. 2010. Paralogismos y sofismas del discurso político español. la falacia política en un corpus de debates parlamentarios. *Anuario de estudios filológicos*.

Sepúlveda-Torres, R., A. Bonet-Jover, I. Diab, I. Guillén-Pacho, I. Cabrera-de Castro, C. Badenes-Olmedo, E. Saquete, M. T. Martín-Valdivia, P. Martínez-Barco, and L. A. Ureña-López. 2024. Overview of flares at iberlef 2024: Fine-grained language-based reliability detection in spanish news. *Procesamiento del lenguaje natural*, 73:369–379.

Tindale, C. W. 2007. *Fallacies and argument appraisal*. Cambridge University Press.

Vallecillo-Rodríguez, M. E., M. V. Cantero-Romero, I. Cabrera-de Castro, L. A. Ureña-López, A. Montejo-Ráez, and M. T. Martín-Valdivia. 2024. Overview of refutes at iberlef 2024: Automatic generation of counter speech in spanish. *Procesamiento del Lenguaje Natural*, 73:449–459.