Case-Based Reasoning with Language Models for Classification of Logical Fallacies

Zhivar Sourati[1,2] , **Filip Ilievski**[1,2] , **Hông-Ân Sandlin**[3] and **Alain Mermoud**[3]
[1]Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA
[2]Department of Computer Science, University of Southern California, Los Angeles, CA, USA
[3]Cyber-Defence Campus, armasuisse Science and Technology, Switzerland
{souratih,ilievski}@isi.edu, {hongan.sandlin,alain.mermoud}@ar.admin.ch

Abstract

The ease and speed of spreading misinformation and propaganda on the Web motivate the need to develop trustworthy technology for detecting fallacies in natural language arguments. However, state-of-the-art language modeling methods exhibit a lack of robustness on tasks like logical fallacy classification that require complex reasoning. In this paper, we propose a Case-Based Reasoning method that classifies new cases of logical fallacy by language-modeling-driven retrieval and adaptation of historical cases. We design four complementary strategies to enrich input representation for our model, based on external information about goals, explanations, counterarguments, and argument structure. Our experiments in in-domain and out-of-domain settings indicate that Case-Based Reasoning improves the accuracy and generalizability of language models. Our ablation studies suggest that representations of similar cases have a strong impact on the model performance, that models perform well with fewer retrieved cases, and that the size of the case database has a negligible effect on the performance. Finally, we dive deeper into the relationship between the properties of the retrieved cases and the model performance.

1 Introduction

The ease and speed of spreading misinformation [Wu *et al.*, 2019; Allcott *et al.*, 2019] and propaganda [Da San Martino *et al.*, 2019; Barrón-Cedeno *et al.*, 2019] on the Web motivate the need to develop trustworthy technology for understanding novel arguments [Lawrence and Reed, 2020]. Inspired by centuries of philosophical theories [Aristotle, 1989; Locke, 1997; Copi, 1954; Barker, 1965], recent work has proposed the natural language processing (NLP) task of *Logical Fallacy Detection*. Logical Fallacy Detection goes beyond prior work on binary detection of misinformation and fake news classification, and aims to classify an argument into one of the dozens of fallacy classes. For instance, the argument *There is definitely a link between depression and drinking alcoholic drinks. I read about it from Wikipedia* belongs to the class *Fallacy of Credibility*, as the validity of the argument is based on the credibility of the source rather than the argument itself. Here, the focus is on informal fallacies that contain incorrect or irrelevant premises, as opposed to formal fallacies, which have an invalid structure [Aristotle, 1989]. The identification of informal fallacies is challenging for both humans and machines as it requires complex reasoning and also common knowledge about the concepts involved in the fallacy [Hansen, 2020]. To predict the correct fallacy type, the model has to know what Wikipedia is and how it is used in societal discourse, the potential relationship between depression and consuming alcoholic beverages, and also the causal link between the first and second parts of the argument.

The currently dominant NLP paradigm of language models (LMs) has been shown to struggle with reasoning over logical fallacies [Jin *et al.*, 2022] and similar tasks that require complex reasoning [Da San Martino *et al.*, 2019; Barrón-Cedeno *et al.*, 2019]. As LMs are black boxes, attempts to improve their performance often focus on adapting their input data. Prior work has pointed to the need to include context [Vijayaraghavan and Vosoughi, 2022], simplify the input structure [Jin *et al.*, 2022], or perform special training that considers soft logic [Clark *et al.*, 2021]. However, these ideas have not been successful in classifying logical fallacies yet. Alternatively, methods that leverage reasoning by example, e.g., based on Case-Based Reasoning (CBR), have shown promise in terms of accuracy and explainability for other tasks like question answering [Das *et al.*, 2022], but have not been applied to reason over logical fallacies to date. We conclude that integrating such explainable methods with generalizable LMs provides an unexplored opportunity to reason over logical fallacies.

In this paper, we pursue the question: *Does reasoning over examples improve the ability of language models to classify logical fallacies?* To answer this question, we develop a method based on the idea of CBR [Aamodt and Plaza, 1994]. We focus on the interpretive problem-solving variant of CBR, which aims to understand novel cases in terms of previous similar cases while not necessarily using the solutions from previous cases directly [Leake, 2001]. We adapt this idea to the task of classifying logical fallacies by using LMs as backbones when retrieving and adapting prior similar cases. We measure the ability of our models in terms of accuracy and generalizability and also probe their explainability. The main contributions of this paper are as follows:

1. We design the first **Case-Based Reasoning method** for logical fallacy classification to solve new cases based on past similar cases. The framework implements the theory of CBR with state-of-the-art (SOTA) techniques based on language modeling and self-attention.
2. We design four **enriched case representations**: *Counterarguments*, *Goals*, *Explanations*, and *Structure* of the argument to allow CBR to retrieve and exploit similar cases based on implicit information, like argument goals. To our knowledge, we are the first who investigate the effect of these case representations on CBR performance.
3. We perform **extensive experiments** that investigate the impact of CBR against Transformer LM baselines on in-domain and out-of-domain settings. We perform ablations to provide insight into the sensitivity of our CBR method on its parameters and investigate the explanations extracted from the model.

We make our code and data available to support future research on logical fallacy classification.[1] For additional discussion, please refer to longer version of the paper on Arxiv.

2 Method

CBR [Schank, 1983] is a method that reasons over new cases based on similar past cases with a known label [Aamodt and Plaza, 1994]. Our CBR formulation (Figure 1) consists of three steps: (1) given a new case, *retrieve* similar cases from a case database, (2) *adapt* fetched similar cases based on the current one, and (3) *classify* the new case based on the adapted exemplars. In this work, we use LMs in the retriever and the adapter because of their strong ability to encode and compute similarity for any textual information.

Retriever. Finding k similar cases $S_i \quad (i \in \{1, ..., k\})$ to the new case C from a case database is retriever's task. The retriever estimates the similarity between C and S_i by encoding each of them with the same LM encoder and computing the cosine similarity of the resulting encodings. The retriever then picks the k cases with top cosine similarities from the database. The new case is concatenated to its similar cases, i.e., $S = C \oplus < SEP > \oplus S_1 \oplus S_2 \oplus ... \oplus S_k$ and is passed as input to the CBR adapter.

Adapter. The framework's middle part aims to prioritize the most relevant information from S for reasoning over the new case C. Based on the second step of the pipeline by [Aamodt and Plaza, 1994], after fetching similar cases, it might be the case that only certain retrieved cases would be useful, and therefore, they should be weighted according to their utility for approaching the new case. The fusion of the current case with its previously seen similar problems would give the model the chance to come up with a better representation of the current problem, as well as better abstractions and generalizations for further uses. The adapter consists of two parts: an *encoder* and an *attention component*. The encoder is an LM that takes as an input C and S separately, then outputs their respective embedding representations E_C and E_S. We use the hidden states of the last layer of the LM as the input embedding. A multi-headed attention component [Vaswani *et al.*, 2017] with H heads selects the most useful information from the similar cases embeddings E_S given the embedding of the new case E_C. As commonly done in Transformer architectures, the Adapter generates Value and Key vectors from E_S and Query vectors from E_C. The dot product of the Query and Key vectors, fed through a softmax layer, results in an Attention vector, which indicates the importance of each token in S when generating the adapted vector A. An adapted vector with adjusted attention on its elements is produced by the weighted sum of the Value vectors based on Attention weights. The output of the attention component is A, the adjusted embedding of E_S.

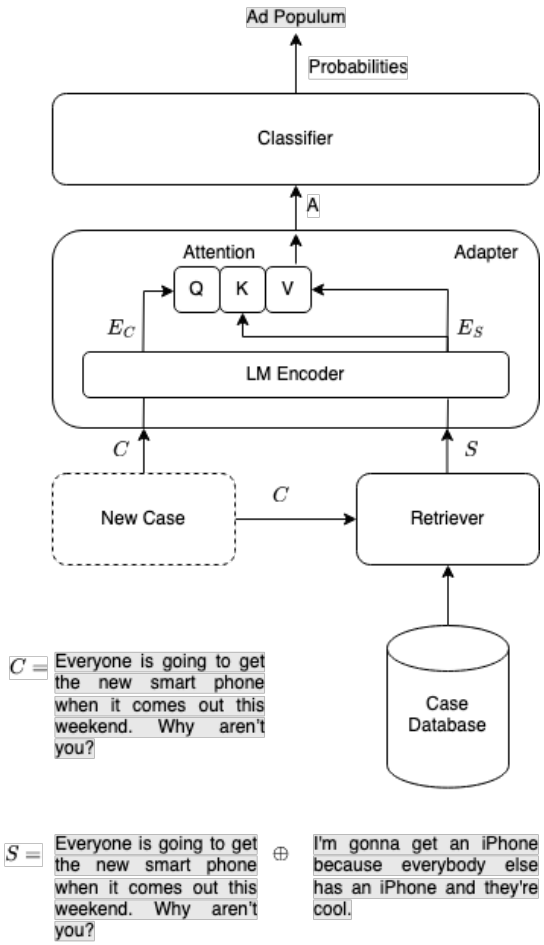

Figure 1: Three stages of the CBR pipeline. Using the new case C, the retriever finds k similar cases $\{S_1, S_2, ..., S_k\}$ and creates $S = C \oplus < SEP > \oplus S_1 \oplus S_2 \oplus ... \oplus S_k$. The adapter processes both the new case and fetched similar cases and tries to adapt S based on the new case C, and extracts more abstract information from the fusion of the two. Finally, the classifier receives the adapted information and returns the probabilities associated with the new class belonging to each fallacy type. In the example, $k = 1$.

Classifier. Last part of the framework predicts the final class based on the adapter output A. The classifier is de-

[1] https://github.com/zhpinkman/CBR

signed as a fully connected neural layer with a depth d and an activation function. The objective function of the classifier is the cross-entropy loss. The cross-entropy loss is computed over the probabilities that are extracted from C logits corresponding to each of the C classes. Also, during training, the retriever's component weights are frozen while the rest of the framework is trained in an end-to-end fashion.

Overall, our CBR architecture resembles a standard 'vanilla' LM with a classification head but brings the additional benefit of having access to prior relevant labeled cases weighed based on the attention mechanism.[2] We hypothesize that the CBR models bring two benefits over vanilla LMs: (1) the integration of similar labeled cases helps the model analyze the new fallacious argument better and classify it more accurately, and (2) provides explicit insights into the reasoning of the model by yielding similar cases to the current one [Renkl, 2014].

3 Case Representation

Merely retrieving labeled cases may not be sufficient for reasoning on new cases, as it is unclear what dimensions of similarity their relevance is based on. For instance, two cases may be similar in terms of their explanation, structure, or the goal behind the cases. As these dimensions are implicit and not apparent from the plain text, we make them explicit by enriching the original text of the case with such information. We consider four representations in which the case formulation is enriched with its *counterargument*, *goal*, *explanation*, and *structure*. As a baseline, we also include the original *text* without any enrichments. Table 1 illustrates examples of these representations for the sample case *There was a thunderstorm with rain therefore I did not finish my homework*.

Each of the enrichment strategies r modifies the case representation by concatenating it with additional information, $r(case)$. We introduce a case representation function $R(case, r)$ that concatenates $case$ with additional information $r(case)$ resulting in $case \oplus r(case)$. These representations modify both the new case C to $R(C, r)$ and cases from the database S_i to $R(S_i, r)$, and change the cosine similarity to be computed between enriched cases instead of plain text. We next describe the design of the enrichment strategies.

Counterarguments. Counterarguments are common in persuasive writing, where they explain why one's position is stronger than the counterargument and serve as a preemptive action to anticipate and remove any doubts about arguments [Harvey, 2009]. We hypothesize that counterarguments are often implicit in the arguments, and would therefore be useful to be provided directly to the model. For instance, in the argument presented in Table 1, although the plain text claims that the reason for not finishing the homework is the heavy rain, the counterargument points out *other reasons for not finishing the homework such as the person being too tired*.

Goals. Studies of argumentation often focus on the interplay between the goals that the writer is pursuing and their argumentations [Tracy, 2013]. Thus, when classifying logical fallacies, we expect that it is beneficial to take into account the goals of the arguments. The goal may be entirely missing in the argument's text, or the argument may implicitly hint at the goal. An example of the latter is shown in Table 1, where the phrase *therefore I did not finish my homework* alludes to the implicit goal of the writer to *justify not finishing their homework*. As shown in this example, we include an explicit goal statement to fill this gap.

Explanations. By using explanations about logically fallacious arguments, we aim to augment the arguments with a broader notion of information that might be useful for classifying logical fallacies but is not already included in the original argument, such as reasoning steps getting from premises to conclusions of an argument [Barker, 1965]. As we do not impose any restrictions on the explanations, their content may overlap with the previous two representations. Alternatively, explanations may provide different complementary information. Such is the example in Table 1 that discusses the *causal relationship between two events that are not actually related*. Thus, the explanation acts as a general gap-filling mechanism that can provide any relevant information that is missing in the original argument.

Structure. Tasks like logical fallacy classification involve higher-order relation comprehension that is often based on the structure rather than the content of the argument. In that sense, the semantics of specific entities and concepts in the argument may be misleading to the model. Similarly to [Jin *et al.*, 2022], we hypothesize that focusing on the logical structure of an argument rather than its content is beneficial for the model's performance [Gabbay *et al.*, 2004]. An example of a structural simplification of an argument is presented in Table 1. While this simplification may help the model grasp the case structure more directly, the structure formulation may not detect the implicit causal links between the thunderstorm (X) and the homework (Z).

We extract the enrichment information for a case using a combination of few-shot and zero-shot prompting with two SOTA models: ChatGPT [OpenAI, 2022] and Codex [Chen *et al.*, 2021]. Given a representation strategy r, we prompt ChatGPT to get the representations for a case for five different examples using one template per representation. For instance, we use the template *Express the goal of the argument* $\{case\}$ to retrieve the *goals* of the argument $case$. The five obtained examples per representation are used as demonstrations to prompt Codex in a few-shot manner. For a representation strategy r, we use the same demonstrations together with each new case C from our task as input to the Codex model, which yields enrichment information $r(case)$ per case. In this manner, we combine the strong zero-shot ability of the closed-source ChatGPT model with the few-shot generation strength of the Codex model.

4 Experimental Setup

In this section, we describe the evaluation data and metrics, the baselines we compare to, and the implementation details.

Evaluation dataset. We use two logical fallacy datasets from [Jin *et al.*, 2022], called LOGIC and LOGIC Climate.

[2] Our experiments using the framework without the attention mechanism consistently showed sub-optimal performance.

Representation	Transformed Text
Goals	It's possible that the goal is to explain why the speaker did not finish their homework. The speaker may be trying to convince the listener that they did not finish their homework because of the thunderstorm.
Counterarg.	There are many factors that contribute to a person's ability to complete their homework, and it's not fair to suggest that the thunderstorm was the only factor. It's possible that the person did not finish their homework because they were distracted by the thunderstorm or because they were tired.
Explanations	It presents a causal relationship between two events that might not be actually related.
Structure	There was an X with Y therefore I did not do Z.

Table 1: One example of different representations for the case *There was a thunderstorm with rain therefore I did not finish my homework*.

The LOGIC dataset includes thirteen logical fallacy types about common topics, namely: *Ad Hominem, Ad Populum, Appeal to Emotion, Circular Reasoning, Equivocation, Fallacy of Credibility, Fallacy of Extension, Fallacy of Logic, Fallacy of Relevance, False Causality, False Dilemma, Faulty Generalization,* and *Intentional*. LOGIC Climate dataset consists of more challenging examples for the same logical fallacy types on the climate change topic. We use LOGIC for in-domain evaluation and LOGIC Climate for out-of-domain evaluation. As LOGIC dataset is severely imbalanced, we augment its train split using two techniques, i.e., back-translation, and substitution of entities in the arguments with their synonymous terms. This augmentation makes LOGIC dataset train split have 281 arguments for each fallacy type. Note that we do not fine-tune our model on the LOGIC Climate dataset in our experiments to evaluate the generalizability of our framework. We model the classification task as a multi-class classification problem and use the customary metrics of weighted precision, recall, and F1-score.

Baselines. We consider three different LMs: BERT [Devlin *et al.*, 2018], RoBERTa [Liu *et al.*, 2019], and ELECTRA [Clark *et al.*, 2020]. We apply our CBR method (§2) on each of these models. As baselines, we use vanilla LMs without a CBR extension. We also compare against Codex in a few-shot setting, with the prompt including all the possible classes as well as one example for each class resulting in thirteen labeled examples in the prompt. Finally, we include the results of a frequency-based predictor that predicts fallacy classes based on the distribution of fallacy types in the training set.

Implementation details. We use SimCSE [Gao *et al.*, 2021], a transformer-based retriever that is optimized for capturing overall sentence similarity, to compute the similarity between cases (§2) and also use $H = 8$ heads for the multi-headed attention component. The depth of our classifier is $d = 2$. It uses $gelu$ [Hendrycks and Gimpel, 2016] as an activation function. We analyze the performance of our model using $k \in \{1, 2, 3, 4, 5\}$. To test the generalization of our model with sparser case databases, we experiment with various ratios of the case database within $\{0.1, 0.4, 0.7, 1.0\}$.

5 Results

In this section, we measure the effectiveness of CBR per model and case representation. We further provide ablations that measure the sensitivity of the model to the size of the case database and the number of cases. Finally, we present a qualitative analysis of the explainability of CBR and a thorough discussion about how retrieved cases help to classify new ones.

		LOGIC			LOGIC Climate		
Model	**Type**	P	R	F1	P	R	F1
Freq-based	baseline	0.094	0.094	0.093	0.120	0.079	0.080
Codex	few-shot	0.594	0.422	0.386	0.198	0.093	0.077
ELECTRA	baseline	0.614	0.602	0.599	0.276	0.229	0.217
	CBR	**0.663**	**0.664**	**0.657**	**0.355**	**0.254**	**0.270**
RoBERTa	baseline	0.577	0.561	0.560	0.237	0.211	0.200
	CBR	**0.631**	**0.619**	**0.619**	**0.379**	**0.248**	**0.245**
BERT	baseline	0.585	0.598	0.586	0.166	0.130	0.120
	CBR	**0.613**	**0.616**	**0.611**	**0.359**	**0.204**	**0.200**

Table 2: Comparison of the best results of the CBR framework with vanilla LMs and two external baselines on two benchmarks focusing on both in-domain (LOGIC) and out-of-domain (LOGIC Climate) settings. The best results per model are **boldfaced** and the overall best results are underlined.

Impact of CBR. Table 2 shows the performance of the CBR framework and relevant baselines. For each model, we present the results using the best case representation per model and using $k = 1$ while exploiting 10% of the case database that we found to yield the best results among all possible combinations. Overall, the CBR method brings a consistent and noticeable quantitative improvement in the classification of logical fallacies by LMs. For each of the three LMs, CBR outperforms the vanilla baselines by 2.5 - 6 absolute F1 points on the in-domain dataset and up to 8 points on the out-of-domain dataset. Furthermore, CBR outperforms Codex, which is utilized in a few-shot setting, despite it being a much larger model. Across the different LMs, ELECTRA is achieving the best score and benefits the most from the CBR framework on the in-domain benchmark, which we attribute to its efficiency of pre-training [Clark *et al.*, 2020]. The same pattern of the superiority of CBR over vanilla LMs can be observed for the other two models with different pre-training procedures and varying numbers of internal parameters. The CBR method notably and consistently improves the performance of the LMs on the out-of-domain (LOGIC Climate) benchmark as well, with ELECTRA performing the best and BERT benefiting the most from CBR.[3] We conclude

[3]Per-class experiments demonstrated the ability of CBR model to improve the accuracy of baseline models for all fallacy types, es-

that CBR is a general framework that can be applied to any LM and can generalize well to unseen data and to various fallacy classes. The generalization of CBR is in line with prior work that suggests its strong performance on tasks with data sparsity [Das *et al.*, 2020].

		LOGIC			LOGIC Climate		
Model	**Representation**	P	R	F1	P	R	F1
ELECTRA	*Text*	0.655	0.634	0.635	0.317	0.242	0.242
	Counterarg.	**0.663**	**0.664**	**0.657**	0.355	**0.254**	**0.270**
	Goals	0.646	0.622	0.621	**0.376**	0.217	0.222
	Structure	0.634	0.625	0.618	0.375	0.254	0.269
	Explanations	0.605	0.580	0.578	0.314	0.242	0.237
RoBERTa	Text	**0.633**	0.613	0.619	0.343	0.236	0.251
	Counterarg.	0.624	0.613	0.615	0.367	0.198	0.216
	Goals	0.632	0.613	0.619	0.351	0.242	**0.263**
	Structure	0.631	**0.619**	**0.619**	**0.379**	**0.248**	0.245
	Explanations	0.575	0.558	0.559	0.359	0.192	0.181
BERT	*Text*	0.595	0.604	0.596	0.311	0.192	0.204
	Counterarg.	0.607	0.613	0.603	0.342	**0.217**	**0.228**
	Goals	0.598	0.607	0.596	0.310	0.204	0.203
	Structure	**0.613**	**0.616**	**0.611**	**0.359**	0.204	0.200
	Explanations	0.540	0.531	0.532	0.274	0.217	0.190

Table 3: Performance of the CBR framework using different case representations. The best results per model are **boldfaced** and the overall best results are underlined.

Effect of different representations. The results in Table 3 confirm our expectation that the case representation plays an important role in the effectiveness of the CBR framework. Depending on the LM used, the performance difference among different case representations ranges from 6 to 8% F1-scores for the in-domain setting and 4 to 8% F1-scores for the out-of-domain setting. In general, we observe a boost in performance when enhancing the original representation (*text*). *Counterargument* information yields the highest boost, though the impact of the representations varies across models. Using ELECTRA, the enrichment with *counterarguments* helps the most, outperforming the model based on the original *text* and the other enrichment strategies. With RoBERTa, *goals* and *structure* of the arguments perform on par with *text*, while with BERT, including information about *counterarguments* and argument *structure* outperforms the *text* representation. As the LMs have been trained with different data and may optimize for different notions of similarity, it is intuitive that the impact of the case representations varies across models. This finding is in line with theoretical work, which discusses that knowledge transfer is strictly guided by the similarity function of the reasoning model [Holyoak and Thagard, 1996]. Meanwhile, using a generic enrichment with *explanations* performs consistently poorly and harms the model performance, which suggests that the CBR models benefit from more precise case representations.

Effect of case database size. Next, we investigate the sensitivity of the best-performing CBR model based on ELECTRA to the size of the case database. Figure 2 (left) depicts the performance of this model using different ratios of the case database. The figure shows that the CBR framework consistently outperforms the vanilla LM baseline (with 0% of cases) on in- and out-of-domain settings. This trend stands regardless of the size of the case database, which indicates the low sensitivity of the CBR model to the case database size. However, we note that using 10% of the case database yields the best performance, which indicates that a limited case database offers a better potential of abstraction to CBR. Moreover, comparing the performance of the model using different ratios of the case database, we observe a continuous decrease in the performance using higher percentages of the case database. Having access to too much data makes the model dependent and sensitive to the unnecessary and insignificant details of similar cases retrieved. These observations point us to the data efficiency properties of the CBR framework [Das *et al.*, 2020].

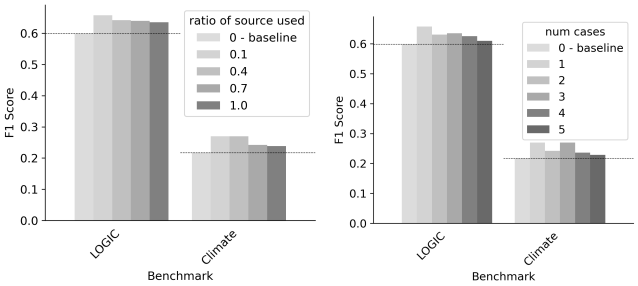

Figure 2: Performance of the CBR framework using different ratios of the case database (left) and different numbers of cases (right). The baseline is outlined as the dotted line.

Effect of different number of cases. The performance of the best CBR model that uses ELECTRA with different numbers of cases is illustrated in Figure 2 (right). In both in-domain and out-of-domain settings, we observe a consistent pattern of performance decrease when more cases are taken into account in the reasoning process. The CBR framework reaches its peak performance using only one similar case while once more outperforming the vanilla LM (with 0 cases) in all the settings. This indicates that the models get easily overwhelmed with past information when considering a new case. While intuitively, one would expect that a larger number of cases should help the model analyze a new case better, the reasoner should have the capacity to process all of these past cases. Otherwise, as observed in this experiment, including more cases can have an adverse effect on the reasoner by distracting it rather than helping it.

Case study on explainability. A key promise of the CBR framework is its native explainability by cases since its retrieval of similar cases and reasoning over them are integrated into the CBR process. We perform a qualitative analysis of the cases retrieved by CBR to develop a better intuition about its reasoning process. Table 4 illustrates four example cases that the vanilla LM classifies incorrectly. For each case, we show two CBR representations: one leading to a correct prediction and one leading to an incorrect one. The first example shows the scenario where the original *text* of a retrieved case does not suffice for the model to reason correctly, despite its topical surface similarity to the input case. In other words, the high surface similarity of similar cases is confusing the model and forcing it to incorrectly predict the same class that

pecially the ones having the least number of training examples.

Input Sentence	Enriched Representation for Correct Prediction (representation)	Enriched Representation for Wrong Prediction (representation) (predicted class)	Class
People who don't support the proposed minimum wage increase hate the poor.	There are often multiple perspectives on an issue. It's possible to have a nuanced or balanced view that doesn't align with any side completely. ***(Counterarg.)***	That candidate wants to raise the minimum wage, but they aren't even smart enough to run a business. ***(Text) (Ad Hominem)***	*Fallacy of Extension*
The house is white; therefore it must be big.	X is y; therefore, it is z. ***(Structure)***	The sentence "People who drive big cars hate the environment" presents a generalization about a group of people without sufficient evidence and it relies on oversimplification.***(Explanations) (Faulty Generalization)***	*Fallacy of Logic*
Student: You didn't teach us this; we never learned this. Teacher: So, you're either lazy or unwilling to learn is that right?	It's possible that the argument "It's possible to pass the class without attending. so, you will pass even if you don't attend" is trying to convince the listener that they will pass the class even if they don't attend. The speaker may be trying to persuade the listener to skip class. ***(Goals)***	The sentence "Teacher: You are receiving a zero because you didn't do your homework. Students: Are you serious? You gave me a zero because you hate me?" attacks the person making the argument rather than the argument itself. ***(Explanations) (Fallacy of Extension)***	*False Dilemma*
One day, Megan wore a Donald Duck shirt, and she got an A on her test. Now she wears that shirt every day to class.	There are many factors that contribute to a student's grade, and it's not fair to suggest that the student's past grades are the only factor. It's possible that the student failed the test because they didn't study, or because they were sick. ***(Counterarg.)***	The sentence "Eating five candy bars and drinking two sodas before a test helps me get better grades. I did that and got an A on my last test in history" presents a causal relationship between two events without sufficient evidence to support the claim. ***(Explanations) (Fallacy of Relevance)***	*False Causality*

Table 4: Four examples from different classes in which the CBR model predicts the correct class. For each example, we show a representation that leads to a correct prediction and a representation that still leads to predicting the wrong class. We also show the corresponding wrong class predicted by the second variation of the model.

is associated with the retrieved similar case. However, we see that enriching the case with its *counterargument* helps the CBR model, even though the counterargument is phrased in an abstract manner and is not similar to the new case on the surface. We observe a similar situation with the *explanations* enrichment in the third example, having high surface similarity between the retrieved case and the new one, where analyzing the argument *goals* instead helps the model. In the second example, the *structure* of the argument and the logical depiction of the past cases help the most, while in the fourth example, the *counterarguments* assist the reasoning of CBR. From the second and the fourth example, we observe that enriching arguments with cases that are semantically far from the new case is confusing for the CBR model, even if their reasoning would be helpful.

Representation	LOGIC		LOGIC Climate	
	ground truth overlap	predictions overlap	ground truth overlap	predictions overlap
Text	0.184	0.232	**0.136**	0.173
Counterarg.	0.208	0.220	0.062	0.068
Goals	0.178	0.196	0.130	0.124
Structure	0.238	0.250	0.105	0.242
Explanations	**0.277**	**0.447**	0.086	**0.478**

Table 5: Overlap of retrieved cases' labels with true labels and predictions of the best CBR model (ELECTRA). We highlight the highest overlaps in **bold**.

In summary, presented examples show that the retrieved cases help the model indirectly by providing CBR with high-level information (first example), symbolic abstractions (second example), extensive analysis of the writer's goal (third example), and alternative possibilities (fourth example). This brings up a natural question: does CBR performance correlate to class overlap between the current case and retrieved similar cases? In other words, can we label a new case solely based on its k-nearest neighbors' labels? To answer this question, we compute the overlap of retrieved cases' labels with both the true and the predicted label for different case representations (Table 5). We observe a low overlap of a maximum of 27.7% between the retrieved cases' labels and the true labels, which is only slightly better than a frequency-based prediction. Also, centering on the direct effect of retrieved cases on the CBR predictions, the model with the highest class overlap between the retrieved cases and the predicted classes also has the lowest performance (*explanations*). Meanwhile, the best CBR variants (e.g., *counterarguments*) do not directly reuse the labels of the retrieved cases. We conclude that while retrieving similar cases provides the CBR models with useful information, this additional evidence influences the model reasoning indirectly and may have adverse effects otherwise. Although CBR, in its simplest form, can act as a k-nearest neighbors algorithm, our results suggest that the neighbors' labels cannot be used blindly, and further reasoning step over the retrieved cases is necessary. We believe that these findings open exciting future research directions that investigate the relationship between case similarity and CBR performance.

6 Related Work

In this section, we present prior research on logical fallacy classification, CBR, and methods that prompt very large LMs.

Logical fallacy. Prior computational work on logical fallacies has mostly focused on formal fallacies using rule-based systems and theoretical frameworks [Nakpih and Santini, 2020]. Nevertheless, recent work has switched attention to informal logical fallacies and natural language input. Jin *et al.* propose the task of logical fallacy classification, considering thirteen informal fallacy types and two benchmarks. The authors gather a rich set of arguments containing various logical fallacies from online resources and evaluate the capabilities of large LMs in classifying logical fallacies both in in-domain and out-of-domain settings. Similarly, Goffredo *et al.* present a dataset of political debates from U.S. Presidential Campaigns and use it to evaluate Transformer LMs. Processing different parts of arguments, such as dialogue's context, they create separate expert models for each part of arguments and train all the models together, from which they report the importance of discussion context in argument understanding. Although LMs have been used to classify logical fallacies, both independently and in an ensemble setting, to our knowledge, no prior work has tried to improve LMs' capabilities to reason over previous cases of logical fallacies encountering a new case nor experimented with enriching the argument representation. We fill this gap by employing CBR with LMs to reason over similar past cases to classify logical fallacies in new cases.

Case-Based Reasoning. Case-Based Reasoning [Schank, 1983] has been a cornerstone of interpretable models in many areas. For instance, researchers have applied CBR over past experiences in mechanical engineering [Qin and Regli, 2003] and medical applications [Oyelade and Ezugwu, 2020]. Case-Based Reasoning has been also used in education, particularly to teach students to recognize fallacies [Spensberger *et al.*, 2022]. Exploiting its interpretable properties, Walia *et al.* use Case-Based Reasoning as a transparent model for Word Sense Disambiguation, Brüninghaus and Ashley use Case-Based Reasoning for predicting legal cases an interpretable pipeline, while Ford *et al.* use Case-Based Reasoning to enhance the transparency of classifications made on written digits [Lecun *et al.*, 1998]. Inspired by its advantages, we couple Case-Based Reasoning with LMs, leading to enhanced accuracy and explainability of classifying logical fallacies. To our knowledge, this is the first work that combines CBR with LMs for complex tasks like logical fallacy classification. Nevertheless, there are frameworks that are close to CBR that also have a notion of memory, but cannot serve as replacements, given their restrictions. Analogical reasoning [Gentner and Smith, 2012] methods typically focus on proportions between words or short text sequences and cannot generalize well to unstructured text. K-nearest neighbor methods are a simplified version of CBR that, given our observations, can not perform as well as CBR. Our framework can also be seen broadly as a memory-based model [Weston *et al.*, 2014], however, our proposed formulation that combines CBR and LMs has not been explored before for tasks like logical fallacy classification.

Prompting LMs. The behavior of LMs is dependent on the quality of their inputs. Aiming to create more comprehensive inputs for LMs and assist them in complex reasoning tasks, researchers have attempted to transfer knowledge from very large LMs to smaller ones. Shwartz *et al.* show that LMs can discover useful contextual information about the question they answer, from another LM. Wang *et al.* propose an LM pipeline that learns to faithfully reason over prompt-based extracted rationales. Wei *et al.* explore how generating a series of intermediate reasoning steps using prompting can equip LMs with complex reasoning skills. Inspired by the ability of large LMs to provide relevant information for novel inputs, as well as prior work that performs knowledge distillation from large to smaller LMs [West *et al.*, 2021], we use prompting to enrich the arguments containing logical fallacies. According to [Barker, 1965], logical fallacies are created by transition gaps from premises to conclusions, and we try to enrich the arguments using prompting to cover the gaps. Our method resembles retrieval-augmentation methods [Lewis *et al.*, 2020], yet, our enrichment strategies are novel and have not been explored on such complex tasks.

7 Conclusions and Future Work

In this paper, we presented a novel method that uses Case-Based Reasoning with LMs to classify logical fallacies. The CBR method reasons over new cases by utilizing past experiences. To do so, the method retrieves the most relevant past cases, adapts them to meet the needs of a new case, and finally classifies the new case using the adjusted information from past cases. We devised four auxiliary case representations that enrich the cases with implicit information about their counterarguments, goals, structure, and explanations. Our results showed that CBR can classify logical fallacies and can leverage past experiences to fill the gaps in LMs. CBR outperformed the LM baselines in all settings and across all thirteen logical fallacy classes. CBR was able to generalize well and transfer its knowledge to out-of-domain setting. The representation of its cases played a key role: enriching cases with counterarguments helped the most, while adding generic explanations harmed the model's performance. Furthermore, CBR models performed best when a small number of cases are provided, but showed low sensitivity to the size of the case database. Finally, our qualitative analysis demonstrated the value of CBR as an interpretable framework that benefits from past similar cases indirectly.

Since our experiments showed that similar cases assist CBR indirectly, future research should further qualify the relationship between the information provided by the retrieved cases and the performance of the model. Moreover, future work should focus on evaluating CBR on other natural language tasks that require abstraction, such as propaganda detection and dialogue modeling. For instance, given a task-oriented dialogue about cooking a new meal, the model may benefit from procedures for cooking similar meals. The application of CBR on such tasks might also inspire additional case enrichment strategies, e.g., that describe the causal relation between text chunks, and point to additional knowledge gaps that CBR needs to fill.

Acknowledgements

Zhivar Sourati has been supported by armasuisse Science and Technology, Switzerland under contract No. 8003532866, and NSF under Contract No. IIS-2153546, while Filip Ilievski is sponsored by the DARPA MCS program under Contract No. N660011924033 with the US Office Of Naval Research.

References

[Aamodt and Plaza, 1994] Agnar Aamodt and Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, 7:39–59, 1994. 1.

[Allcott *et al.*, 2019] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2):2053168019848554, 2019.

[Aristotle, 1989] Aristotle. *On sophistical refutations: On Comin to be passing away - on the cosmos v. 3*. Loeb Classical Library. LOEB, London, England, July 1989.

[Barker, 1965] Stephen Francis Barker. *The Elements of Logic*. New York: Mcgraw-Hill, 1965.

[Barrón-Cedeno *et al.*, 2019] Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864, 2019.

[Brüninghaus and Ashley, 2006] Stefanie Brüninghaus and Kevin D Ashley. Progress in textual case-based reasoning: predicting the outcome of legal cases from text. In *AAAI*, pages 1577–1580, 2006.

[Chen *et al.*, 2021] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[Clark *et al.*, 2020] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*, 2020.

[Clark *et al.*, 2021] Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021.

[Copi, 1954] Irving M. Copi. Introduction to logic. *Philosophy*, 29(110):271–271, 1954.

[Da San Martino *et al.*, 2019] Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 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, 2019.

[Das *et al.*, 2020] Rajarshi Das, Ameya Godbole, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. A simple approach to case-based reasoning in knowledge bases. In *Automated Knowledge Base Construction*, 2020.

[Das *et al.*, 2022] Rajarshi Das, Ameya Godbole, Ankita Naik, Elliot Tower, Manzil Zaheer, Hannaneh Hajishirzi, Robin Jia, and Andrew McCallum. Knowledge base question answering by case-based reasoning over subgraphs. In *International Conference on Machine Learning*, pages 4777–4793. PMLR, 2022.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Ford *et al.*, 2020] Courtney Ford, Eoin M. Kenny, and Mark T. Keane. Play mnist for me! user studies on the effects of post-hoc, example-based explanations &; error rates on debugging a deep learning, black-box classifier. *arXiv preprint arXiv:2009.06349*, 2020.

[Gabbay *et al.*, 2004] Dov M Gabbay, John Hayden Woods, et al. *Handbook of the History of Logic*, volume 2009. Elsevier North-Holland, 2004.

[Gao *et al.*, 2021] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[Gentner and Smith, 2012] D. Gentner and L. Smith. Analogical reasoning. In V.S. Ramachandran, editor, *Encyclopedia of Human Behavior (Second Edition)*, pages 130–136. Academic Press, San Diego, second edition edition, 2012.

[Goffredo *et al.*, 2022] Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. Fallacious argument classification in political debates. In *Thirty-First International Joint Conference on Artificial Intelligence {IJCAI-22}*, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization, 2022.

[Hansen, 2020] Hans Hansen. Fallacies. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2020 edition, 2020.

[Harvey, 2009] Gordon Harvey. A brief guide to the elements of the academic essay. *Harvard College Writing Program*, 2009.

[Hendrycks and Gimpel, 2016] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[Holyoak and Thagard, 1996] Keith J Holyoak and Paul Thagard. *Mental leaps: Analogy in creative thought*. MIT press, 1996.

[Jin *et al.*, 2022] Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[Lawrence and Reed, 2020] John Lawrence and Chris Reed. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818, 01 2020.

[Leake, 2001] D.B. Leake. Problem solving and reasoning: Case-based. In Neil J. Smelser and Paul B. Baltes, editors, *International Encyclopedia of the Social & Behavioral Sciences*, pages 12117–12120. Pergamon, Oxford, 2001.

[Lecun *et al.*, 1998] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Locke, 1997] John Locke. *An Essay Concerning Human Understanding*. Penguin classics. Penguin Classics, London, England, June 1997.

[Nakpih and Santini, 2020] Callistus Ireneous Nakpih and Simone Santini. Automated discovery of logical fallacies in legal argumentation. *International Journal of Artificial Intelligence & Applications*, 11(2):37–48, mar 2020.

[OpenAI, 2022] OpenAI. Chatgpt. https://openai.com/blog/chatgpt, 2022. Accessed: April 30, 2023.

[Oyelade and Ezugwu, 2020] Olaide N. Oyelade and Absalom E. Ezugwu. A case-based reasoning framework for early detection and diagnosis of novel coronavirus. *Informatics in Medicine Unlocked*, 20:100395, 2020.

[Qin and Regli, 2003] Xiaoli Qin and William C. Regli. A study in applying case-based reasoning to engineering design: Mechanical bearing design. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 17(3):235–252, 2003.

[Renkl, 2014] Alexander Renkl. Toward an instructionally oriented theory of example-based learning. *Cognitive Science*, 38(1):1–37, 2014.

[Schank, 1983] Roger C. Schank. *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*. Cambridge University Press, USA, 1983.

[Shwartz *et al.*, 2020] Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unsupervised commonsense question answering with self-talk. *arXiv preprint arXiv:2004.05483*, 2020.

[Spensberger *et al.*, 2022] Florian Spensberger, Ingo Kollar, and Sabine Pankofer. Effects of worked examples and external scripts on fallacy recognition skills: a randomized controlled trial. *Journal of Social Work Education*, 58(4):622–639, 2022.

[Tracy, 2013] Karen Tracy. *Understanding face-to-face interaction: Issues linking goals and discourse*. Routledge, 2013.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[Vijayaraghavan and Vosoughi, 2022] Prashanth Vijayaraghavan and Soroush Vosoughi. TWEETSPIN: Fine-grained Propaganda Detection in Social Media Using Multi-View Representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3433–3448, Seattle, United States, July 2022. Association for Computational Linguistics.

[Walia *et al.*, 2019] Himdweep Walia, Ajay Rana, and Vineet Kansal. Case based interpretation model for word sense disambiguation in gurmukhi. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 359–364, 2019.

[Wang *et al.*, 2022] PeiFeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. PINTO: Faithful language reasoning using prompt-generated rationales. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*, 2022.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[West *et al.*, 2021] Peter West, Chandrasekhar Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. Symbolic knowledge distillation: from general language models to commonsense models. In *North American Chapter of the Association for Computational Linguistics*, 2021.

[Weston *et al.*, 2014] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

[Wu *et al.*, 2019] Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. Misinformation in social media: Definition, manipulation, and detection. *SIGKDD Explor. Newsl.*, 21(2):80–90, nov 2019.