LogicJitter: Let LLMs play Logic Games and they will Detect Misinformation

Anonymous authors
Paper under double-blind review

Abstract

In the face of the growing challenge of information overload online, the ability to accurately distinguish between genuine information and misinformation has become increasingly critical both from an individual and from a societal point of view. Methodologies for misinformation detection predominantly rely on supervised approaches, which depend heavily on large labeled datasets. However, these datasets are not only costly and time-consuming to produce, but they are also susceptible to issues such as labeling bias, time leakage, the inherent subjectivity of the task, and domain-specific limitations. In this paper, we aim to overcome the aforementioned challenges by proposing a novel and cost-effective strategy to enhance the logical reasoning capabilities of Large Language Models (LLMs), thereby improving their ability to detect misinformation. Our approach, termed LogicJitter, employs a data augmentation technique during fine-tuning that generates both correct and incorrect statements within rule-based logic games. These games are designed to counteract well-known human cognitive biases and logical fallacies. Hence, the primary contributions of this work include demonstrating the effectiveness of logical reasoning fine-tuning on LLMs and providing an open source package for the automatic generation of correct and incorrect logic-based training data, to ease reproducibility. Experimental results confirm this approach improves misinformation detection.

1 Introduction

In an increasingly connected and automated world, where information is generated and disseminated online at an unprecedented speed and volume, often without reliable oversight, the risk of *information manipulation* grows daily (Wardle & Derakhshan, 2017; Wu et al., 2019). As a result, distinguishing between authentic information and misinformation has become more crucial than ever. Indeed, falling into misinformation can lead to serious problems both for individuals and for society as a whole. For example, individuals may make harmful decisions about their health by following false medical advice. On a broader scale, misinformation can fuel distrust in public institutions, spread false narratives, and even incite social unrest, destabilizing communities (OECD, 2024).

Most of the approaches used in recent years in the literature for misinformation detection rely on *supervised learning* solutions (i.e., classifying information as true or false) (Hu et al., 2022b; Reis et al., 2019; Viviani & Pasi, 2017). However, these approaches come with several challenges. First, they require large amounts of labeled data, which is often costly and time-consuming to obtain (Reis et al., 2019). Second, they struggle to adapt to new types of misinformation, as the specific datasets they are trained on may be domain-dependent and not generalize well to other domains (Clarke et al., 2020; D'Ulizia et al., 2021). Lastly, these models can oversimplify the issue, as misinformation is not always clearly true or false but may exist in gray areas of partial truths or misleading contexts (Cabitza et al., 2022).

Recently, there has been growing interest in using *Large Language Models* (LLMs) to address the problem of misinformation detection (Hu et al., 2024; Papageorgiou et al., 2024). However, several open challenges remain. One key issue is that LLMs can sometimes generate or amplify misinformation themselves, as they rely on vast amounts of data from various sources, including unreliable ones, possibly leading to *factual inconsidency* and *hallucinations* (Ji et al., 2023). Furthermore,

LLMs can also be prone to biases present in their training data, which can affect the accuracy and fairness of their outputs. Lastly, there are concerns about the *interpretability* and *transparency* of these models (Zhao et al., 2024), making it difficult to understand how they reach their conclusions.

It is in this context that this article proposes a novel approach to circumvent the scarcity of labeled data for misinformation detection with LLMs. We hypothesize that enhancing the *logical reasoning* abilities of an LLM can significantly improve its capability to detect misinformation. In fact, misinformation often exploits human *cognitive biases* and *fallacies* (French et al., 2023; Stanovich, 2003), which we hypothesize could be more easily identified if an LLM is trained to focus on *logical consistency* (Dillehay et al., 1966; Kainz, 1995). Therefore, we propose and evaluate the fine-tuning of LLMs using carefully designed datasets that encourage logical reasoning, while simultaneously fine-tuning them for misinformation detection. Since purely natural language datasets are often ambiguous and domain-dependent, we propose training the models on *logic games* that are algorithmically generated but can be readily explained in natural language. In this logic games, proofs are presented either with or without errors, and with distractors that model situations where cognitive biases and fallacies would come into play. The LLM is then trained to detect flaws in the reasoning. The objective is to strengthen the model's logical rigor, thereby enhancing its ability to identify common human biases and, consequently, misinformation. We therefore focus on addressing the following two research questions:

RQ1. Can LLM's ability to reason logically in texts be improved with rule-based logic games?

RQ2. Will LLMs trained to detect cognitive biases and logical fallacies help in addressing the misinformation detection task?

Our proposed approach, namely *LogicJitter*, is a form of data augmentation for textual datasets, named in analogy to *ColorJitter* (Zini et al., 2023), a data augmentation technique for image datasets. In LogicJitter, true and false statements based on logic games are interspersed within the training batches to introduce variability and improve the model's reasoning capabilities.

Building on the previous premises, our key contributions are as follows:

- We demonstrate that fine-tuning LLMs to identify logical errors in structured logic games, a method we term LogicJitter, effectively enhances their ability to discern between accurate information and misinformation;
- To ease reproducibility, we provide both a PyTorch and a HuggingFace package that automatically generates correct and incorrect logic games, with the flexibility to be extended with additional logic games.

2 Related Work

Training LLMs to solve Rule-Based Problems. The use of synthetic data generation has been investigated as a means of providing more data in language modeling and reasoning tasks. For instance, (Gunasekar et al., 2023) leverage LLM-generated data with notable performance on reasoning-based tasks. However, when LLMs are trained on rule-based problems, they are usually tested on the same rule-based problems. Generating synthetic geometric data for pre-training and fine-tuning the model on auxiliary constructions to address specific theorems was used to tackle Olympiad-level math problems (Trinh et al., 2024). However, LLMs seem to generalize worse in out-of-distribution data, assessed by length generalization, compared to graph networks trained on the same algorithmically generated problems (Veličković et al., 2022; Markeeva et al., 2024). Formal languages have been used to understand if LLM truly were equally capable of learning languages that are possible and impossible for humans to learn (Kallini et al., 2024). It would be however valuable to be able to leverage a rule-based training, to generalize on natural language datasets.

LLMs at Causal and Logical Reasoning. LLMs have been shown to be able to learn to detect structures in complex causal networks, when trained on simple ones (Vashishtha et al., 2024). Complex self learning loops have can allow LLMs to improve their ability to reason by coming up with their own rationales (Zelikman et al., 2022). Despite their progress, LLMs still struggle with complex mathematics and external computation libraries and symbolic solvers are becoming standard to

produce LLMs that are stronger in maths (Gou et al., 2024). Formal benchmarks exist to assess the causal reasoning abilities of LLMs (Jin et al., 2023; 2024), focusing on the identification of direct and indirect causal relationships. Their findings underscore the shortcomings of LLMs in executing accurate causal reasoning.

Cognitive Biases and Fallacies. Datasets and tasks to detect logical fallacies are available, but usually rely on human annotators that have to classify the errors encounter as a type of fallacy or as another (Jin et al., 2022). On the other hand, AI systems have been observed to reproduce some human cognitive biases such as the confirmation bias, the primacy effect, the representativeness bias, the anchoring bias, and problems related with causality (Martínez et al., 2022). Contrary to recent studies showcasing advanced reasoning abilities in LLMs, randomized controlled trials demonstrate the existence of anchoring bias in all models tested (Nguyen, 2024).

LLMs and Misinformation Detection. It has been shown that capabilities of LLMs to mislead humans are already superior to those of humans to mislead humans, and that the best detectors of LLM lies are LLMs themselves (Chen & Shu, 2024). However, developing datasets for fact checking is often costly given that it can require experts in the field of interest (Kotonya & Toni, 2020a). Therefore, datasets for misinformation detection are not only rare but also small (Schlichtkrull et al., 2024).

3 METHODOLOGY

In this section we define the cognitive biases and fallacies we are interested in compensating for (under the hypothesis that helping LLMs detect them will help them detect misinformation), we give the details of LogicJitter and the logic games we use in this work, we detail the considered LLMs, and we discuss the fine-tuning procedure.

3.1 COGNITIVE BIASES

Human cognitive biases are psychological tendencies that often lead to unconsciously distorted thinking and decision-making. They can be generally regrouped into (Van Eyghen, 2022; Gigerenzer, 2002): (i) *belief, decision-making and behavioral*, (ii) *social*, and (iii) *memory* biases. Of the long list of well-studied biases, we describe those we try to compensate for with *LogicJitter*.

Belief, decision-making, and behavioral biases. Many biases relate to the unexpected speed and direction of updates of beliefs in human decisions. In fact, it has been shown that humans update their beliefs slower than Bayes rule, possibly a consequence of having to deal with noise in the memory recall and in the evidence acquisition processes (Hilbert, 2012). Therefore a slow belief update could be Bayes optimal in the presence of memory noise or mistrust in the sources for example, which could also be modeled as noise in the evidence acquisition process. Some biases emphasize belief updates in long-term memories, that could be stored at the synapse level, and some in short-term memories, possibly stored as spiking activity. The cognitive biases we think can better be compensated algorithmically are the tendency to revise beliefs insufficiently when presented with new evidence (*conservatism bias*), the tendency to reject evidence that contradicts established norms (*Semmelweis reflex*), to ignore the general prevalence, in favor of the information pertaining only to a specific case (*base rate fallacy*), to expect a member of a group to have certain characteristics without having actual information about that individual (*stereotyping*), to misinterpret statistical experiments involving conditional probabilities (*Berkson's paradox*) and the tendency to fail to recognize that a plan of action is no longer appropriate for a changing situation (*plan continuation bias*).

Social biases. Our thinking about other people often follows distorted patterns, leading us to overestimate, underestimate, or misjudge them. Some examples are the tendency to trust more the opinion of an authority figure, regardless of the content (*authority bias*), for people to seem more attractive in a group than alone (*cheerleading effect*), for someone's positive or negative traits to influence how others perceive them in unrelated areas (*halo effect*), the tendency to believe that physically attractive individuals also have intelligence, good judgment, or other positive personality traits (*physical attractiveness*) or the tendency to do and believe as others (*bandwagon effect*).

Memory biases. Some cognitive biases can enhance or hinder memory recall or can distort the content of the recalled memory. Some examples that we tackle by LogicJitter are the tendency to prefer easily available examples (*availability bias*), the fact that unusual or strange information is remembered more effectively than ordinary information (*bizarreness effect*) or the tendency to recall better items that are at the beginning or at the end of a list (*primacy and recency effects*).

3.2 FALLACIES

Fallacies are errors in logical reasoning that affect the validity of arguments, which can be either intentional or unintentional. While both cognitive biases and fallacies can lead to incorrect conclusions, cognitive biases are more about how we think, and fallacies are about how we argue.

Formal fallacies. They are errors in the argument's logical form in a formal logical system. One example is *illicit commutativity* which takes the form: 'If A then B. Therefore, if B then A'. Another example is *denying the antecedent*, such as: 'If A then B. Therefore, if not A then not B'. Neither of the two is generally true, which is why they are considered fallacies;

Informal fallacies. They are false because they rely on false premises. An example of an informal fallacy that we saw above as a cognitive bias is the *authority bias*, where an argument is deemed true only on the premise of the position of the person making the statement.

3.3 LOGICJITTER

To account for the data scarcity in misinformation detection and to compensate for the aforementioned forms of cognitive biases and fallacies, we introduce a data augmentation technique that we name *LogicJitter*, based on *logic games*, described textually, and *distractors* that make reasoning about them more difficult, but still possible in an exact manner.

3.3.1 LOGIC GAMES

We generated algorithmically the following games that are used in the fine-tuning phase to increase the reasoning ability of LLMs.

Guided Maths. This is a dataset with step-by-step solutions to mathematical equations. We use the *scratchpad* technique (Nye et al., 2021; Zelikman et al., 2022) to show the steps necessary to solve three types of sub-problems: *addition*, *multiplication*, and *polynomial evaluation*. Whenever we introduce an *error*, we make sure the mathematical proof we provide in the scratchpad assumes the mistake was correct and builds on top of it.

Causal Clauses. A dataset where we generate complex graphs of causal links, and ask if a statement is true within that graph. Each graph has between three and six nodes, one-tenth of the time the net will be linear, and the rest it will be equally likely Erdos-Renyi, Watts-Strogatz or Barabasi-Albert (Albert & Barabási, 2002), with randomized edge direction. The two subproblems we propose are: (i) determining if two random nodes are connected or not, and (ii) determining if a random node is a fork, a collider, none, or both.

Context-Free Grammars. We build random context-free grammars (Hoperoft & Ullman, 1979), that are not recurrent and have a maximum of 5 non-terminals and 4 terminals. Then we design two sub-problems: (i) provided a generation, the LLM is asked if it belongs to the grammar, and (ii) provided with up to 4 grammars, the LLM is asked which one can produce the most or the least amount of sentences.

CLEVR and CLEAR. We leverage datasets like CLEVR (Johnson et al., 2017), for visual reasoning and CLEAR (Abdelnour et al., 2018; 2023), for acoustic question-answering, to provide varied logical challenges, in this case spatial and temporal logical challenges. We use up to 8 templates, as sub-problems, to generate a diverse set of complex questions about the scenes.

3.3.2 DISTRACTORS

We previously introduced the first two distractors, i.e., the *diversity of games* and related *sub-problems*, and the *diversity of errors* we introduced in half of the generations.

Errors. Usually, datasets algorithmically generated are created without mistakes, and human-generated textual datasets are likely to have errors but it is difficult to know if this is the case. Here we introduce mistakes algorithmically, to always know where they are. We do it because we want to train LLMs to spot subtle mistakes by themselves, and therefore in half of the samples there is going to be an error.

Random Characters. After presenting the problem, a solution is stated, which may be either correct or incorrect if we introduce an error. Following this, a set of randomly generated characters provide their opinions. We use up to five characters, with an equal probability of any number of them being right or wrong. Consequently, the likelihood of one character being wrong is the same as that of two being wrong, or all being wrong, with all possible outcomes equally likely.

To have a list of characters is thought to provide an opportunity to compensate for the *primacy* and *recency biases*, since the characters providing the correct answer are placed randomly. Also, all characters can be wrong, so the available information might be wrong, providing a chance to compensate for the *availability bias*.

Characters are generated in the form 'one adjective + one noun'. The adjective is picked randomly to describe either a nationality (e.g., Namibian), a similarity description (e.g., like you), a sexual orientation (e.g., bisexual), a religious affiliation (e.g., Buddhist), an ethnic group (e.g., Pacific Islander), a degree of attractiveness (e.g., good-looking), or a character trait (e.g., disrespectful). The noun is picked randomly to describe a family relationship (e.g., cousin), an authority figure (e.g., ambassador), a generic person (e.g., individual), a political orientation (e.g., libertarian), or a group (e.g., alliance). This trick gives a chance to compensate for *stereotyping*, but also for the *bizarreness effect*, given that it encourages a disassociation between personal description and being logically right or wrong. For example, sampling randomly different degrees of attractiveness compensates for the *physical attractiveness bias*, and sampling nouns for groups is intended to compensate for the *bandwagon effect*. Therefore with the random characters we intend to compensate essentially for all the social and memory biases we introduced in Section 3.1.

After presenting the problem and the opinions of the characters, the LLM is asked to provide if one of them was right or wrong, and this time the correct answer is given without error.

Problem Revision. After that, a modification in the initial problem statement is provided, such as new connections in the causal net, or objects removed from the CLEVR scene. The same characters appear again, to provide an opinion on the problem, and they are assigned again randomly a correct or an incorrect answer, independently from the first round. The revision is designed to compensate for the belief biases, and make the LLM take into account new evidence to revise beliefs against the *conservativism bias*, or changing the truths that condition the replies against *Berkson's paradox*, for example. Also revising the initial problem statement is a 'not A' statement, A being the initial problem, therefore providing by default 'not B' as an answer would be incorrect, which is designed to compensate for the *denying the antecendent* fallacy. We did not target directly the *illicit commutativity*, given that logically, the statement 'if B then A' is true if and only if 'not A then not B' is true, and we assume compensating *denying the antecendent* will automatically compensate for *illicit commutativity*.

The convenience of this approach is, among other things, that it should not be subject to labeling bias, time leakage, the inherent subjectivity of the task nor domain-specific limitations. For example in the case of time leakage, LogicJitter is not dependent on fine-tuning information that is potentially in the future of the misinformation that we are trying to classify. For those focused on explainability, it would be relatively straightforward to generate an explanation for the occurrence of an error within a logic game. However, given the already extensive nature of our problem descriptions, we opted not to incorporate the explainability component to avoid further complexity.

Examples of the logic games and distractors contained in LogicJitter are illustrated in Table 1. In green are highlighted the random characters that compensate for social and memory biases, that could be used stereotypically by LLMs. Given that being right or wrong is assigned randomly to the characters in the games, our goal is to train the model not to use that information to evaluate their answer, and only evaluate them within the context of the game. In blue, is highlighted the problem revision, designed to modify the content of the initial problem, and ask the LLM to reevaluate the new answers of the characters, and to reevaluate its own understanding of the scene, to compensate for belief biases and fallacies.

Table 1: LogicJitter presents textual logic games with errors and distractors, but the truth value remains exact. We highlight in green the random characters that compensate for social and memory biases, that could be used stereotypically by LLMs. We highlight in blue the problem revision, designed to compensate for belief biases and fallacies.

Logic Games
Guided maths Input: $7x^3$ for $x = 7$ Target: $< scratch >$, $7x^3 = 7 * (7)^3 = (7) * (343) = 770$, 770, $< /scratch >$, 770. A quaint crew says it's fine. Is the quaint crew correct? False. At a second try it is shown that Input: $7x^3$ for $x = 7$ Target: $< scratch > 7x^3 = 7 * (7)^3 = (7) * (343) = 2401$, 2401, $< /scratch >$, 2365. A quaint crew says it's not ok. Is the quaint crew correct? True.
Causal Clauses Visualize that A fixes B, B fixes C, D fixes A, D fixes C. For this reason, C fixes A. A clique from your country says it's correct, a woman from your region says it's wrong, a socialist from another region says there's no error, a queer club says it's fine, a queer crew says it's not good. Is the woman from your region correct? True. It was later brought to the attention that A does not fix B. Hence C fixes A. A socialist from another region says C doesn't fix A, a woman from your region says it's not correct, a queer crew says it's not ok, a clique from your country says it's right, a queer club says C doesn't fix A. Is the queer crew correct? True.
Context-Free Grammars Given grammar 0, [...], grammar 1, [...], grammar 2, [...] Which grammar produces the largest number of sentences? Grammar 2. A heterosexual provost says it's not correct, a native american liberal says it's not correct, a pansexual community says grammar 1. Is the pansexual community correct? False. Grammar 0 was changed for [...] Which grammar produces the smallest number of sentences? Grammar 1. A pansexual community says it's not correct, a native american liberal says it's not correct, a heterosexual provost says there is an error. Is the heterosexual provost correct? True.
CLEVR and CLEAR There is a very large metal tourmaline tetrahedron at (-0.44, -1.46), a small glass aquamarine calendar at (-0.51, 1.03), a small amber gray remote control at (-1.67, -1.44), a small amber tourmaline remote control at (-1.32, -0.87), a very large amber apatite printer at (1.45, -1.46), and a small amber labradorite pen at (-1.08, -1.18). Is the number of labradorite amber pens right of the glass calendar greater than the number of tiny tourmaline remote controls that are in front of the tourmaline amber remote control? no. A sikh brother says it's fine, a sister from a different city says it's correct. Is the sister from a different city correct? True. The last object has been removed. Are there more small tourmaline amber things right of the tourmaline remote control than gray things in front of the very large amber printer? no. A sikh brother says yes, a sister from a different city says there's no error. Is the sister from a different city correct? True.

3.4 Fine-Tuning Models

We fine-tune two pre-trained language models to enhance their logical reasoning and misinformation detection capabilities: i) GPT-2 (small), a model with 125M parameters (Radford et al., 2019), fine-tuned using Low-Rank Adaptation of Large Language Models (LoRA) (Hu et al., 2022a), a Parameter-Efficient Fine-Tuning (PEFT) technique; ii) LlaMA 3.2 (small), a model with 1B parameters (Meta Llama Team, 2024), fine-tuned with AdaLoRA (Zhang et al., 2023), which dynamically adjusts parameter importance during fine-tuning. Both models are trained using a language modeling objective (next-step prediction with cross-entropy loss) and early stopping based on validation loss. To address misinformation detection, we introduce an additional binary classification loss, weighted ten times more heavily than the language modeling loss. This binary loss determines whether the model can correctly classify a sentence as truthful or not. For input, we use a prompt like: "Is the previous text truthful? Reply with True or False." We then assess whether the predicted probability for the *True* token exceeds that of the *False* token. To further improve robustness, we add noise to

the embedding vectors using NEFTune (Jain et al., 2024), a technique designed to enhance model performance by introducing controlled perturbations.

4 Experimental Evaluation

In this section, we present the experimental evaluation conducted to assess the effectiveness of the proposed data augmentation approach w.r.t. the considered misinformation detection task. We begin by introducing the datasets used in the evaluation, followed by a discussion of the results obtained.

4.1 Misinformation Datasets

In order to improve the logical abilities of LLMs, we propose augmenting a *Target Task Dataset* (TTD) with data generated using LogicJitter. We evaluate our approach on three datasets used as TTDs: *PubHealth* (Kotonya & Toni, 2020b), *VitaminC* (Schuster et al., 2021) and *ISOT* (Ahmed et al., 2017). *PubHealth* comprises 11,832 claims for fact-checking across a wide range of health-related topics, including biomedical subjects and government healthcare policies. *VitaminC* is a multi-domain fact-verification dataset based on Wikipedia edits, containing 488,904 data points. *ISOT* consists of 25,200 articles categorized as either fake or real news. The truthful articles were obtained by crawling *Reuters.com* while the fake news were collected from websites flagged by *Politifact*. To evaluate the effectiveness of our data augmentation technique, we compare it against two baselines: fine-tuning directly on the TTD and augmenting the TTD using an existing human-labeled misinformation dataset, VitaminC in our case. For consistency, we limit the augmented TTD size to no more than twice the original TTD size. In our pipeline, we first augment the dataset, for example ISOT, by adding an extra amount of data points generated using either LogicJitter or VitaminC. The resulting dataset is then shuffled to ensure randomness before training the model.

Both PubHealth and VitaminC provide claims along with a corresponding piece of evidence that either supports or rejects the claim for each data point. For a data point consisting of evidence x, claim y, and label z, we format the input to the LLM as: 'Evidence: x. Claim: y. Does the evidence support the claim? Reply with *True* or *False*: z'. This structure ensures clarity for the model and directly addresses the task of evaluating the relationship between the claim and its evidence. In contrast, ISOT provides longer claims x, with corresponding labels z, but without associated evidence. For this dataset, we format the input as: 'x. Is the preceeding text likely truthful and not fake news? reply with *True* or *False*: z'.

4.2 Results

Table 2 illustrates the results in terms of *accuracy* of misinformation detection (i.e., classifying claims into true and fake). As it emerges from the table, augmenting the TTD both with LogicJitter and with human-labeled data, such as VitaminC, improves the validation and the test results over the PubHealth dataset. Surprisingly, LogicJitter achieves this without the need for human-labeled data. When the TTD is VitaminC instead, can see that the best performance is achieved with LogicJitter with errors and random characters, but without problem revision.

Table 2: Ablation study to understand which parts of LogicJitter contribute the most. The full LogicJitter is composed by four parts: G stands for game description, E for including generations with errors, C for including random characters, and full stands for GECR, with R for revisions. We compare also to augmenting with the human labelled dataset VitaminC. We show results on the PubHealth and VitaminC datasets using the GPT2 model and fine-tuning with LoRA.

	PubHealth		**VitaminC**	
augmentation	val	test	val	test
TTD	54.8%	54.9%	55.0%	55.5%
TTD + VitaminC	**69.8**%	**68.2**%		
TTD + LogicJitter (G)	42.7%	42.1%	74.5%	74.4%
TTD + LogicJitter (GE)	67.4%	66.2%	74.9%	74.8%
TTD + LogicJitter (GEC)	66.5%	64.8%	**75.0**%	**75.1**%
TTD + LogicJitter (full)	68.1%	66.2%	59.9%	59.7%

We can also observe in Table 3, that both adding and removing data from the PubHealth with the augmentation data were effective in improving generalization of the classification model in terms of test accuracy. Instead, when the TTD were VitaminC and ISOT, and using both GPT2 and Llama, and both LoRA and AdaLoRA, increasing the amount of data was generally better.

Table 3: Test accuracy for different augmentation strategies on the PubHealth, VitaminC and ISOT datasets using GPT2 125M parameters and LLama3.2 1B parameters. We show how much additional data we add to (+) or remove from (−) the TTD with augmentation. Almost any amount of data augmentation improves performance. LogicJitter (LJ) achieves it without the need of human annotated data.

dataset	PubHealth		VitaminC	ISOT
model	GPT2		GPT2	Llama
PEFT	LoRA		LoRA	AdaLoRA
augm.	**+LJ**	**+VitC**	**+LJ**	**+LJ**
+100%	66.2%	68.2%	59.7%	**75.0**%
+75%	68.6%	73.2%	54.8%	
+50%	62.0%	66.5%	65.5%	71.4%
+25%	50.2%	65.1%	**93.0**%	58.1%
TTD	54.9%		55.5%	65.3%
-25%	74.2%	70.4%	53.7%	74.0%
-50%	**79.9**%	69.9%	50.8%	51.9%
-75%	78.9%	74.9%	63.6%	
-100%	55.9%	**79.1**%	51.4%	19.5%

5 Discussion and Conclusions

We have introduced LogicJitter, a data augmentation technique for misinformation detection that is generated algorithmically and therefore does not require human labelled data. We showed that it successfully improved the generalization ability of the model compared to fine-tuning only on the target dataset, or augmenting with existing human labeled data on misinformation. It therefore turns an expensive task, expert labeling, into a cheap task, algorithmic generation. Being able to generate a dataset algorithmically comes with a few convenient factors, such as the fact that is completely balanced in terms of number of true and false statements, and stereotyping biases are completely absent since it is coded in such way. Somewhat ironically, we use the often considered old school rule-based AI, such as context-free grammars and causal networks, to compensate for the shortcomings of the new wave of Deep Learning based AI. We believe our evidence supports **RQ1** and **RQ2** in the positive: rule-based games, inspired by the attempt to compensate for cognitive biases and fallacies, can improve LLMs logical reasoning, shown by their improved ability to detect misinformation.

Another possibility to what we presented would be to use existing human datasets such as GLUE (Wang et al., 2018), and present the wrong label to ask the LLM to estimate the veracity of the answer. We decided however to stick to a purely rule-based approach, to cleanly verify its effectiveness. As a possible future direction, it is interesting to consider fuzzy logic statements, to provide the games with more flexibility to deal with uncertainties. It will also be interesting to attempt at compensating for more cognitive biases and fallacies. Moreover, a rule-based approach such as LogicJitter, could be used to produce an algorithmically generated explanation on why an answer is right or wrong, and could therefore be useful for those interested in explainability.

REFERENCES

Jerome Abdelnour, Giampiero Salvi, and Jean Rouat. CLEAR: A dataset for compositional language and elementary acoustic reasoning. In *Visually Grounded Interaction and Language Workshop (VIGIL)*, 2018.

Jerome Abdelnour, Jean Rouat, and Giampiero Salvi. NAAQA: A neural architecture for acoustic question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4997–5009, 2023.

Hadeer Ahmed, Issa Traore, and Sherif Saad. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1*, pp. 127–138. Springer, 2017.

Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

Federico Cabitza, Davide Ciucci, Gabriella Pasi, and Marco Viviani. Responsible AI in healthcare. *arXiv preprint arXiv:2203.03616*, 2022.

Canyu Chen and Kai Shu. Can LLM-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*, 2024.

Charles LA Clarke, Maria Maistro, Mark D Smucker, and Guido Zuccon. Overview of the TREC 2020 health misinformation track. In *TREC*, 2020.

Ronald C Dillehay, Chester A Insko, and M Brewster Smith. Logical consistency and attitude change. *Journal of Personality and Social Psychology*, 3(6):646, 1966.

Arianna D'Ulizia, Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni. Fake news detection: a survey of evaluation datasets. *PeerJ Computer Science*, 7:e518, 2021.

Aaron M French, Veda C Storey, and Linda Wallace. The impact of cognitive biases on the believability of fake news. *European Journal of Information Systems*, pp. 1–22, 2023.

Gerd Gigerenzer. *Adaptive thinking: Rationality in the real world*. Oxford University Press, 2002.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. ToRA: A tool-integrated reasoning agent for mathematical problem solving. In *The Twelfth International Conference on Learning Representations*, 2024.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

Martin Hilbert. Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. *Psychological bulletin*, 138(2):211, 2012.

JE Hoperoft and JD Ullman. Introduction to automata theory, languages, and computation. addison-wesley publishing company, 1979.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 22105–22113, 2024.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022a.

Linmei Hu, Siqi Wei, Ziwang Zhao, and Bin Wu. Deep learning for fake news detection: A comprehensive survey. *AI open*, 3:133–155, 2022b.

Neel Jain, Ping yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R. Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. NEFTune: Noisy embeddings improve instruction finetuning. In *The Twelfth International Conference on Learning Representations*, 2024.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. Logical fallacy detection. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 7180–7198, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng LYU, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. CLadder: A benchmark to assess causal reasoning capabilities of language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Zhijing Jin, Jiarui Liu, Zhiheng LYU, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation? In *The Twelfth International Conference on Learning Representations*, 2024.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Wolfgang Kainz. Logical consistency. *Elements of spatial data quality*, 202:109–137, 1995.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. Mission: Impossible language models. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

Neema Kotonya and Francesca Toni. Explainable automated fact-checking for public health claims. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7740–7754, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.623.

Neema Kotonya and Francesca Toni. Explainable automated fact-checking for public health claims. *arXiv preprint arXiv:2010.09926*, 2020b.

Larisa Markeeva, Sean McLeish, Borja Ibarz, Wilfried Bounsi, Olga Kozlova, Alex Vitvitskyi, Charles Blundell, Tom Goldstein, Avi Schwarzschild, and Petar Veličković. The clrs-text algorithmic reasoning language benchmark. *arXiv preprint arXiv:2406.04229*, 2024.

Naroa Martínez, Ujué Agudo, and Helena Matute. Human cognitive biases present in artificial intelligence. *Revista Internacional de los Estudios Vascos*, 67(2), 2022.

Meta Llama Team. The Llama 3 Herd of Models. 2024.

Jeremy K Nguyen. Human bias in ai models? anchoring effects and mitigation strategies in large language models. *Journal of Behavioral and Experimental Finance*, 43:100971, 2024.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models, 2021. *URL https://arxiv.org/abs/2112.00114*, 2021.

OECD. *Facts not Fakes: Tackling Disinformation, Strengthening Information Integrity*. 2024.

Eleftheria Papageorgiou, Christos Chronis, Iraklis Varlamis, and Yassine Himeur. A survey on the use of large language models (llms) in fake news. *Future Internet*, 16(8):298, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Julio CS Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81, 2019.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36, 2024.

Tal Schuster, Adam Fisch, and Regina Barzilay. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 624–643, Online, June 2021. Association for Computational Linguistics.

Keith E Stanovich. The fundamental computational biases of human cognition: Heuristics that (sometimes) impair decision making and problem solving. *The psychology of problem solving*, pp. 291–342, 2003.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

Hans Van Eyghen. Cognitive bias: Phylogenesis or ontogenesis? *Frontiers in Psychology*, 13: 892829, 2022.

Aniket Vashishtha, Abhinav Kumar, Abbavaram Gowtham Reddy, Vineeth N Balasubramanian, and Amit Sharma. Teaching transformers causal reasoning through axiomatic training. *ICML Workshop on Large Language Models and Cognition.*, 2024.

Petar Veličković, Adrià Puigdomènech Badia, David Budden, Razvan Pascanu, Andrea Banino, Misha Dashevskiy, Raia Hadsell, and Charles Blundell. The clrs algorithmic reasoning benchmark. In *International Conference on Machine Learning*, pp. 22084–22102. PMLR, 2022.

Marco Viviani and Gabriella Pasi. Credibility in social media: opinions, news, and health information—a survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 7(5): e1209, 2017.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446.

Claire Wardle and Hossein Derakhshan. *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg, 2017.

Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD explorations newsletter*, 21(2):80–90, 2019.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lq62uWRJjiY.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38, 2024.

Simone Zini, Alex Gomez-Villa, Marco Buzzelli, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost van de weijer. Planckian jitter: countering the color-crippling effects of color jitter on self-supervised training. In *The Eleventh International Conference on Learning Representations*, 2023.