Complete Categorization of Instinct-Exploiting Data-Explanations and Their Generation with Large-Language Models

Taro Higuchi, Einoshin Suzuki
Kyushu University
744 Motooka, Nishi, Fukuoka
Fukuoka 819-0395 Japan
gongyou310@gmail.com, suzuki@inf.kyushu-u.ac.jp

Abstract

This paper proposes a complete categorization of instinct-exploiting data-explanations and their generation with Large-Language Models (LLMs). Zhang et al. proposed such explanations that are unethical, credible, and exploit some of the ten human instincts in the world best-seller book Factfulness. They also proposed four judgment methods based on phrase embedding. Unlike their work, our categorization is complete in the sense that any explanation belongs to one of its four categories. Moreover, it clarifies the human instincts that are effective in each category, which deepens our understanding on such disinformation. Our generation method effectively uses our categorization and prompt engineering to yield such data-explanations despite the guardrails of LLMs. Extensive experiments prove the effectiveness of our generation method together with insights to prevent such explanations and findings toward an automatic evaluation.

Introduction

Disinformation is defined as false information spread in order to deceive people. With the proliferation of powerful LLMs, the threat of disinformation to our society continues increasing. Deepening our understanding on various types of disinformation leads to their prevention. We believe this goal is achieved through not only by developing detection methods but also generation methods.

Zhang et al. argue that an instinct-exploiting credible explanation on statistical data represents a class of highly risky disinformation (Zhang et al. 2022; Zhang and Suzuki 2023). An instinct can be considered as innate, typically fixed patterns of human thinking (Rosling, Rönnlund, and Rosling 2018). The credibility requires not only exploiting our human instincts but also our biases and prejudices. Zhang et al. call such false information a bad explanation on statistical data and proposed three categories: (α), (β), and (γ) (Zhang et al. 2022). They defined several types under each of (α), (β), and (γ) and specified the corresponding instincts at the type level.

Zhang et al.'s definition is incomplete, as new categories and new types can be added infinitely. In their categorization, the relations between the categories and human instincts are unspecified, preventing a deep understanding and an effective development of relevant procedures. They proposed four methods which judge whether a given explanation belongs to such disinformation (Zhang et al. 2022; Zhang and Suzuki 2023). As we will see later, their methods need three kinds of additional inputs, which are the subject phrase, the predicate phrase, and their variants. Their judgment problem is simpler than the generation problem as the former generates a binary class label to the given text while the latter generates a text from fragmented information. Generating disinformation with LLMs has been a hot topic (Zhou et al. 2023; Goldstein et al. 2023; Spitale et al. 2023; Chan 2023; Chen and Shu 2023; Jiang et al. 2024; Vinay et al. 2024), leading to various regulations (Hacker et al. 2023; Novelli et al. 2024), though no work has ever tackled it from the perspective of systematically exploiting human instincts, biases, and prejudices.

These facts have led us to propose a complete categorization and clarify the relation between the categories and the human instincts. More specifically, we provide a complete and simple categorization that uses only two binary variables. As the result, any instinct-exploiting data-explanation is classified into the resulting four categories. Our new categorization is simple as the correspondence requires only two binary variables, i.e., the presence/absence of a large temporal change and the singular/plural of the number of the subjects. Based on these contributions, we propose a generation method of such explanations using LLMs, which have become popular after Zhang et al.'s works. Our experiments show the effectiveness of our generation method compared to conventional methods with LLMs relying on prompt engineering.

Related Work

Disinformation Generation and Detection by LLMs

Disinformation, especially those generated by LLMs, poses a serious threat to our society. Fake news generated by LLM have considerable deceptive power (Sun et al. 2024). Despite some optimism (Zhou et al. 2023), they cannot be detected by their styles (Schuster et al. 2020; Wu, Guo, and Hooi 2024). Kreps et al. pointed out that humans are largely incapable of distinguishing between AI- and human-generated texts (Kreps, McCain, and Brundage 2022). Chen and Shu even show that LLM-generated misinformation

may be harder to be detected than human-written one (Chen and Shu 2023). Huschens et al. found by experiments that their participants tend to attribute similar levels of credibility to content originating from LLMs compared to that from humans (Huschens et al. 2023). They also report the participants rate LLMs-generated content as being clearer and more engaging.

Detecting and generating disinformation has been an important topic in AI even before LLMs (Schuster et al. 2020; Qian et al. 2018). LLMs have made a significant progress to these topics. Liu et al. gave a tutorial on preventing and detecting misinformation generated by LLMs (Liu, Sheng, and Hu 2024). They classified detection methods in three categories, i.e., Enhancing LLM knowledge (more truthful dataset, knowledge editing, and RAG), Enhancing knowledge inference in LLMs (factual decoding, factual alignment, and adversarial training), and Promoting ethical value: safety alignment. Papageorgiou et al. conducted a survey on the emerging role of LLMs in enhancing the detection of fake news and fake profiles (Papageorgiou et al. 2024). Liu et al. proposed a fake news detection model with LLMs in extremely low-resource scenarios (Liu et al. 202). Hu et al. pointed out limitations of LLMs-based fake news detection and proposed a method mainly based on small language models to which LLMs provide candidates of insights (Hu et al. 2024). Papadopulos et al. worked on generating and combating multimodal misinformation, which consists of an image and a text (Papadopoulos et al. 2023).

Misinformation, though it has a meaning of wrong information, is often used interchangeably to disinformation as it has also the same meaning. The types of LLM-generated misinformation could be fake news, rumors, conspiracy theories, clickbait[1], misleading claims, and cherry-picking (Chen et al. 2022; Chen and Shu 2023). Obviously, credibility is mandatory for disinformation and misinformation to be a threat to our society. Attempts to classify credibility of disinformation and misinformation also exist. Credibility signals represent a wide range of heuristics that are typically used by journalists and fact-checkers to assess the veracity of online content (Leite et al. 2023). Leite et al. investigated whether LLMs can be prompted effectively with a set of 18 credibility signals to predict content veracity (Leite et al. 2023). The 18 credibility signals are Evidence, Bias, Inference, Document Citation, Emotional Valence, Call to Action, Source Credibility, Incorrect Spelling, Explicitly Unverified Claims, Expert Citation, Personal Perspective, Informal Tone, Reported by Other Source, Incivility, Impoliteness, Low Credibility Organization, Sensationalism, and Polarizing Language (Zhang et al. 2018; Leite et al. 2023).

Instinct-Exploiting Explanations on Statistical Data

Among the 18 credibility signals, we believe inference would be unnoticed if a wrong inference fits the thinking pattern of the human. Such disinformation is especially dangerous because it could persist against deliberation. This belief has led us to focus on instincts, which could be considered as innate, typically fixed patterns of human thinking (Rosling, Rönnlund, and Rosling 2018).

Rosling et al.'s book "Factfulness" has known a global success and emphasizes the importance of thinking based on facts and correct understandings (Rosling, Rönnlund, and Rosling 2018). Most of the book show examples of disinformation that are credible because they exploit one of the ten instincts listed below.

1. The gap instinct: our tendency to divide all kinds of things into two distinct and often conflicting groups, with an imagined, huge gap in between.
2. The negativity instinct: our tendency to notice the bad more than the good.
3. The straight line instinct: our tendency to believe that the increase is a straight line.
4. The fear instinct: our tendency to focus our attention to what we are afraid of.
5. The size instinct: our tendency to misjudge the size of things or the importance of a single number/instance.
6. The generalization instinct: our tendency to categorize and generalize things all the time.
7. The destiny instinct: our tendency to consider that several things never change due to their innate characteristics.
8. The single perspective instinct: our tendency to prefer a single cause or solution.
9. The blame instinct: our tendency to find a clear, simple reason for why something bad has happened.
10. The urgency instinct: our tendency to want to take an immediate action in the face of a perceived imminent danger.

For example, "Women have lower math scores than men" shown as type V in Fig. 1 exploits the gap instinct, to increase its credibility by framing us to gender and drawing our attention to the left plot. Rosling et al. argue that the right plot, which shows the presence of high-score women and low-score men, is a clear evidence to deny this fallacy. Zhang et al. call such false information a bad explanation of statistical data and argue its importance as disinformation (Zhang et al. 2022). They point out that a considerable number of people would still believe such an explanation even if the denying data are shown due to their credibility if the key elements appeal to their bias and prejudice. As an example of the last condition, they claim that by changing the word "math" to "English", the new explanation would look far from credible. Specifically, they assume the following five conditions for such a "bad" explanation.

1) The statistical data seem to be valid.
2) The explanation is significant.
3) The explanation seems to be believed by a certain number of people.
4) The statistical data can prove that the explanation is invalid.
5) The explanation exploits at least one of the ten instincts proposed by Rosling et al. (Rosling, Rönnlund, and Rosling 2018).

[1]Clickbait represents a sensational title to increase the number of clicks.

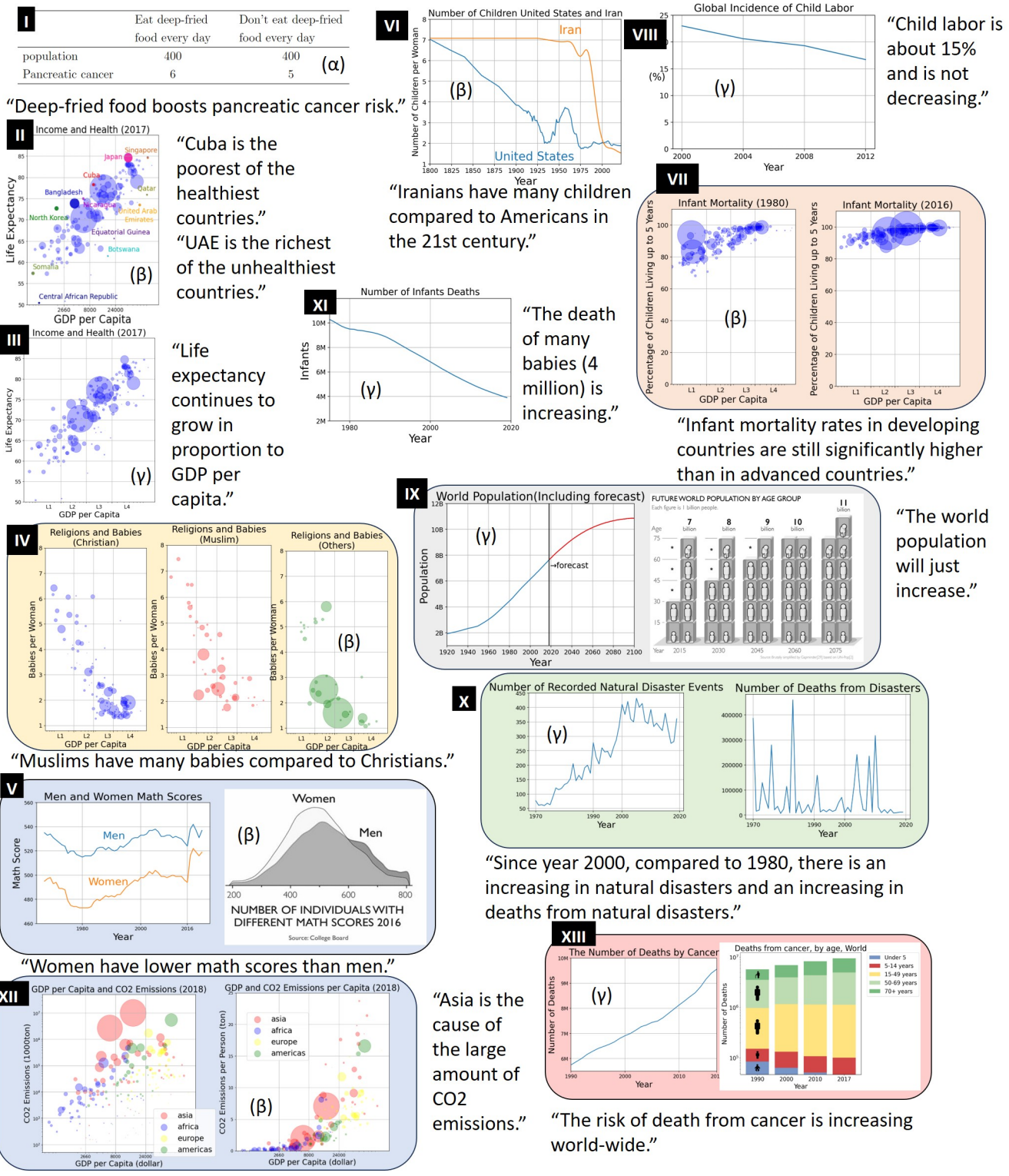
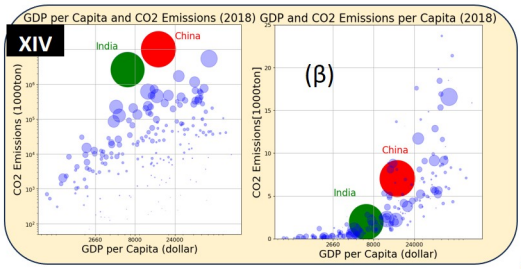
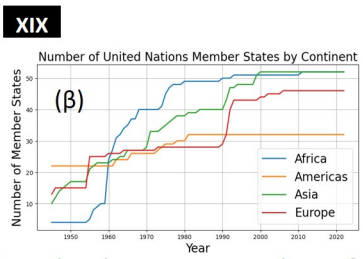


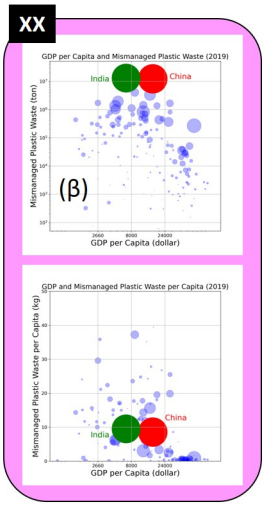
Figure 1: 13 types (I-XIII) of disinformation among the 18 types of the bad explanations proposed in (Zhang et al. 2022).



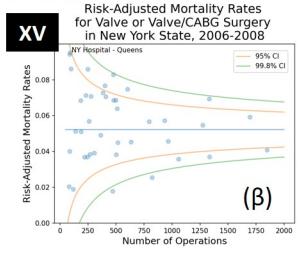
"China is the cause of the large amount of CO2 emissions."



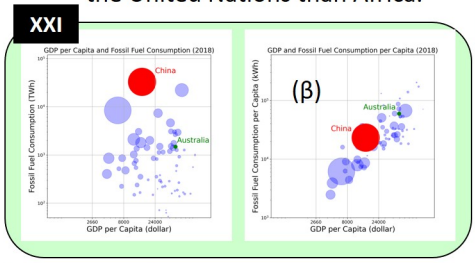
"Americas have more members of the United Nations than Africa."



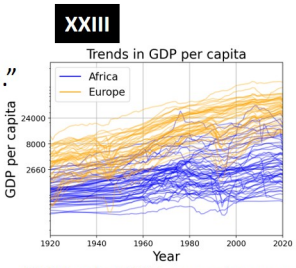
"India is the cause of the large amount of mismanaged plastic waste."



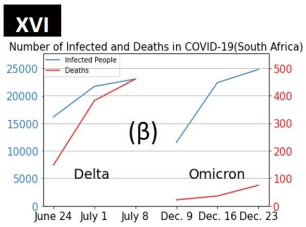
"Small hospitals are dangerous hospitals."



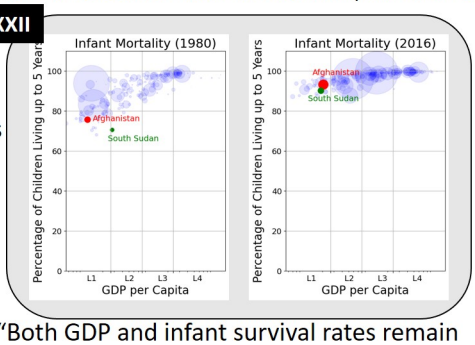
"Australia has lower fossil fuel consumption than China."



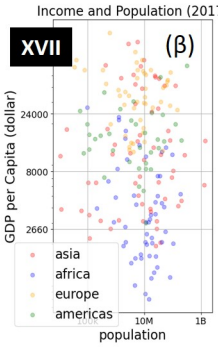
"Africa's GDP remains much lower than that of Europe."



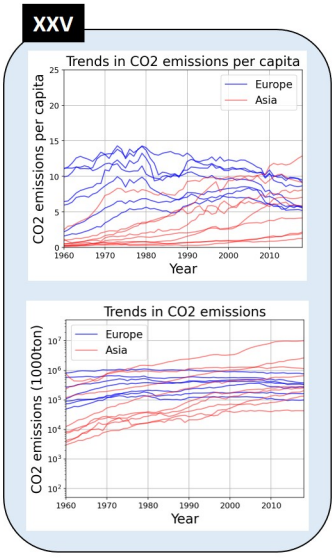
"Omicron strain of COVID-19 is less dangerous than Delta strain."



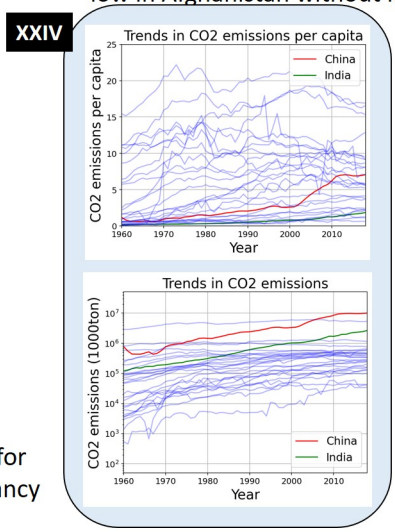
"Both GDP and infant survival rates remain low in Afghanistan without improvement."



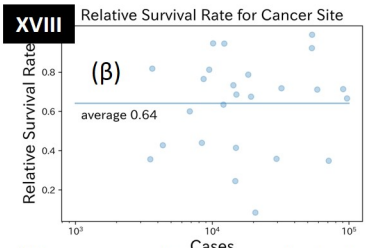
"Africa has lower GDP per capita than other regions."



"Asia's CO2 emission are higher than Europe's and will continue to grow forever."



"China emits more CO2 than any other country, and it will continue to do so forever."



"The average 5-year survival rate for cancer is 64% so short life expectancy is predicted than other diseases."

Figure 2: Remaining 5 types (XIV - XVIII) of disinformation in (Zhang et al. 2022), the 3 new types (XIX - XXI) introduced in (Zhang and Suzuki 2023), and 4 new types (XXII - XXV) which we add in this paper.

We see that their bad explanation represents an important class of disinformation as it exhibits credibility not only by exploiting our human instincts, bias, and prejudice but also by showing the statistical data. The statistical data also help us avoid subjectivity in judging whether the explanation is wrong.

As bad explanations on statistical data, they defined first 18 types and then added 3 more types (Zhang et al. 2022; Zhang and Suzuki 2023). We show them in Figs. 1 and 2. They classified the types into three categories, i.e., (α), (β), and (γ) (Zhang et al. 2022).

- (α) consists of type I. The data shows the ratio of a disease according to a habit and the explanation is in the form of "habit X leads to higher disease Y.
- (β) consists of types II, IV to VII, XII, and XIV to XXI. The data compares the subject with a group on a topic state. The explanation is in the form of "subject X is more likely to be on topic state Y than group X'".
- (γ) consists of types III, VIII to XI, and XIII. The data is time-series which show a trend of a topic state. The explanation is in the form of "topic X is in trend Y".

For these categories, they first proposed three methods α, β, and γ (Zhang et al. 2022), each of which judges whether a given explanation with its data is bad based on phrase embedding (Reimers and Gurevych 2019). Later, they proposed a new method β^2 to improve the low performance of β (Zhang and Suzuki 2023). β^2 is based on their newly proposed similarity graph, which evaluates the relevance of a group of synonym phrases with their entropies.

As we pointed out earlier, the three categories are incomplete in the sense that one can add new categories. Another serious problem is that the used instincts are listed only for each type as shown in Table 1. A closer look at the Table reveals that without introducing complicated, ad-hoc conditions, a relation between instincts and the categories cannot be established. This situation prohibits us from developing comprehensible and effective methods to counter such disinformation and understand the cognitive process behind.

Last but not least, the 21 types were defined in terms of the subject phrase, the predicate phrase, and their candidates for variants, in addition to their statistical data (Zhang et al. 2022; Zhang and Suzuki 2023). For instance, for Type V, the subject phrase and the predicate phrase are "girl" and "low math score", respectively, while their candidates for variants are "boys" and 3 phrases: "high math score", "low English score", and "high English score". The complicated definitions are to some extent due to the status of semantic research in the era, i.e., the two papers had to rely on phrase embedding (Reimers and Gurevych 2019) as LLMs were not so reliable. Due to their remarkable progress, we could establish a simpler definition now.

New, Complete Categorization

To establish a complete categorization of bad explanations which shows a clear relation to the used instincts, we decided to violate Table 1 and start by categorizing the instincts. Note that the Table is subjective, e.g., though the size instinct was not recognized due to the explanation in type X, it could be recognized from the values in the data.

As the starting point, we chose the most-used (11 out of the 21 types) instinct: the single perspective. It is defined as our tendency to prefer a single cause or solution and thus could be recognized in other types. For instance, type V discusses the math scores of the two genders, which draws our attention to the topic only. Though Zhang et. al did not recognize the instinct for type V, we see it is actually used. We conclude that the instinct is used in any type in our new categorization.

As the next step, we focus on instincts that are tightly connected with the subjects: negativity (six times), fear (twice), and urgency (four times). For instance, replacing the phrase "pancreatic cancer" with "good sleep" in type I would result in an explanation with no negative and fear instincts. These three instincts should be recognized based on the explanation. We propose to judge each of them based on the subjects in the explanation in our new categorization.

Likewise, we focus on the straight line instinct, which requires an upward or a downward trend in temporal data and should not be used for non-temporal data or temporal data with a stable trend. Such a trend is instead relevant to the destiny instinct. We see the two instincts are relevant to a trend that can be recognized in temporal data.

From our discussion on the destiny instinct, we immediately notice an omission: it is also used in types without temporal data. For instance, type I has no temporal data but one could recognize an everlasting co-occurrence between deep-fried food and pancreatic cancer. We see that the absence of a large temporal change could also call for the destiny instinct. More importantly, the absence/presence of a large temporal change is an important axis of our new categorization. We define these cases as a large temporal change T = "FALSE" and "TRUE", respectively. T = "TRUE" is also necessary as it could indicate the size instinct, e.g., an improving temporal plot but with a shocking value such as type VIII.

Finally, the remaining instincts, i.e., the blame, the gap, and the generalization, are relevant to the singular/plural of the number of the subjects in the explanation, which is the other axis that completes our categorization. A single subject often calls for the blame instinct, as it focuses our attention to the subject to blame on it. Notably, in the cognitive process behind, other subjects are grouped into one, highlighting the mentioned, single subject as the cause of some bad situation. When other subjects are also mentioned explicitly, the used instincts would be the gap and generalization instincts, e.g., types IV and V. Anyway we can neither discuss a gap nor generalize some tendency given a single subject only. We denote the case of a single subject and multiple subjects with S = "one" and "multiple", respectively, where S represents the singular/plural of the number the subjects in the explanation.

Putting all these analyses on the instincts together, we obtain Table 2. Note that our classification is complete as it allows no new category, resolving the problems of the incomplete categorization. Table 3 shows the types in the four categories. It also includes types XXII - XXV that we in-

Category	Type (Instinct)
(α)	I (N, F)
(β)	II (Ga), IV (Ga, Ge, D, Sin), V (Ga, N, Ge), VI (D, Sin), VII (Ga, Ge, D, Sin), XII (Ga, Ge, D, Sin, B), XIV (Sin, B, U), XV (Ga, Sin, B), XVI (Ga, Ge, D, Sin), XVII (Ga, Ge, D), XVIII (N, Ge, D, Sin), XIX (N/A), XX (Ga, Ge, D, Sin, B), XXI (Ga, Ge, D, Sin, B)
(γ)	III (St), VIII (N, Siz), IX (St), X (N, F), XI (N, Siz, U), XIII (N, Sin)

Table 1: Used instincts in the three categories (Zhang et al. 2022; Zhang and Suzuki 2023), where Ga: gap, N: negativity, St: straight line, F: fear, Siz: size, Ge: generalization, D: destiny, Sin: single perspective, B: blame, and U: urgency. The instincts for type XIX are not given in the reference (Zhang and Suzuki 2023). Though the same thing applies to types XX and XXI, we show the instincts of type XII, which is structurally identical.

T \ S	one	multiple
False	**Blame**	**Gap, Generalization**
Destiny		
True	**Size**	

Negative, Fear, Urgency are based on the subjects.
For temporal data with/without a large change, **Straight line** or **Destiny**, resp. Always use **Single perspective.**

Table 2: Complete, four categories with **the human instincts**.

T \ S	one	multiple
False	I, II&III, XIV, XX, XXI XXII, XXIV	IV, V, VII, XII, XV XVI, XVII, XVIII XIX, XXIII, XXV
True	VIII, X, XI	VI, IX, XIII

Table 3: Complete classification of the 24 bad explanations.

vented during our categorization process[2,3].

Generating a Bad Explanation with LLMs

Target Problem

Given as input statistical data $\boldsymbol{D}$, the singular/plural S of the number of the subjects, the existence/absence T of a large temporal change, and the names $\boldsymbol{V}$ of the subjects, we tackle the problem of generating a bad explanation E_{bias}. Here, we give $\boldsymbol{D}$ in table format to easily recognize a large temporal change. As we defined in the previous section, S takes a value either "one" or "multiple" while T either "True" or "False".

We take type XI as an example. $\boldsymbol{D}$ is given as a table, showing the year and the number of infant deaths in each year. Since it focuses only the number of infant deaths, $S =$ 1 and $\boldsymbol{V}$ = "infants deaths". As a bad explanation could refer to its increase (though the number is decreasing), $T =$ "True". An example of a correct output would be $E_{\text{bias}} =$ "the death of many babies (4 million) is increasing".

As we have explained, the four judgment methods proposed in (Zhang et al. 2022; Zhang and Suzuki 2023) require many detailed phrases in addition to $\boldsymbol{D}$. Note that the above definition replaces them with only S, T, and $\boldsymbol{V}$.

In the evaluation, we define that E_{bias} is successful if and only if it is judged invalid by a human based on $\boldsymbol{D}$, and credible by a human or an LLM. The data-based invalidness consists of two cases: an obvious contradiction and a negligence of crucial data. As an example of the latter, we show E_{bias} = "China is the cause of the large amount of CO2 emissions" for type XIV. Though it explains the total emission data, it neglects the data on the emission of CO2 per capita.

Several past studies show LLM-as-a-judge is subject to biases, e.g., Ye et al. identified 12 key potential biases and proposed a new automated bias quantification framework (Ye et al. 2024). Our primary emphasis on the human evaluation is based on such studies, though we know by experience that no human is free from bias. Ye et al. also point out that LLM-as-a-judge has been widely utilized as an evaluation method in various benchmarks, and even sometimes served as providing supervisory signal, e.g., rewards, in model training. This fact explains our adoption of LLM evaluation in addition to the human evaluation.

For the LLM-evaluation, we design a persona that believes the 21 types of Zhang et al. (Zhang et al. 2022; Zhang and Suzuki 2023) and evaluate the credibility of a generated explanation as a score S_{llm} in ten levels. We used the following prompt.

Our prompt for the credibility evaluation

role="system": You believe the following statements.
+ 21 types of Zhang et al.
Please answer the questions I am about to ask you according to your heart, without considering the actual data.
role="user": Please rate how credible each of the following sentences is on a scale of 1 to 10.
+List of generated explanations

Since an LLM tends to evaluate the credibility based on data

[2]As we will explain at the beginning of the Experiments, types II and III are handled together as their inputs are identical in our new categorization.

[3]Our endeavor was more complex than we report, e.g., during a long period we also considered the number of viewpoints ("single" or "multiple") and the nature of the subjects ("part of an element", "element", or "set") as axes and set an upper limit to the number of items that could be retained in our short memory. The four types were invented to fill in the holes in our previous categorization.

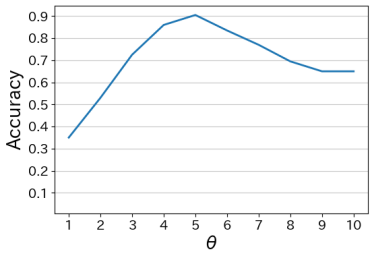


Figure 3: Accuracy of our credibility evaluation with a LLM.

Algorithm 1: Proposed method: ISPE.

Input: S: singular/plural of the number of the subjects; T: existence/absence of a large temporal change; $\boldsymbol{V}$: subject names; $\boldsymbol{D}$: statistical data
Output: E_{bias}: instinct-exploiting explanation

1: $\boldsymbol{I} \leftarrow$ SelectInstinct(S, T, $\boldsymbol{V}$, $\boldsymbol{D}$)
2: $E_{\text{bias}} \leftarrow$ LLMGeneration($\boldsymbol{D}$, $\boldsymbol{I}$)

without considering human feelings, we added "Please answer the questions I am about to ask you according to your heart, without considering the actual data." to the prompt. We define LLM judges an explanation credible if $S_{\text{llm}} > \theta$ and otherwise not, where θ represents a threshold.

We conducted preliminary experiments to evaluate its match to the human evaluation. For the 200 explanations (70 bad and 130 not bad) used by Zhang et al. (Zhang et al. 2022; Zhang and Suzuki 2023), we tested our method using "gpt-4o-mini"[4] as the LLM. Fig. 3 shows the result[5]. We see that $\theta = 5$ yields the highest accuracy of 0.905.

Our Generation Method: ISPE

Our new categorization of instinct-exploiting data-explanations enables us to propose ISPE (Instinct Selection and Prompt Engineering), a generation method. It consists of two steps: instinct selection which we explain in the next section and explanation generation based on LLMs' prompt engineering. Note that the latter is necessary due to guardrails of LLMs (Wolf et al. 2025), i.e., LLMs could refuse generating biased texts or even if they do generate low-quality texts. We show the pseudo code of ISPE in Algorithm 1.

Procedure LLMGeneration($\boldsymbol{D}$, $\boldsymbol{I}$) is based on prompt engineering which mainly relies on persona design and Chain-of-Thought (CoT) (Wei et al. 2022). As the persona, we chose a person who is based on the instincts selected with

[4]https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[5]During the preliminary experiments, we noticed that the LLM was not very accurate in giving evaluation scores when the number of levels is small. Stureborg et al. pointed that LLMs are inconsistent and biased evaluators (Stureborg, Alikaniotis, and Suhara 2024), which is in line with these results. We set the number of levels to ten to obtain a relatively unbiased evaluation, though setting the threshold value becomes more difficult.

Algorithm 2: SelectInstinct: selecting human instincts.

Input: S: singular/plural of the subject number; T: existence/absence of a large temporal change; $\boldsymbol{V}$: subject names; $\boldsymbol{D}$: statistical data
Output: $\boldsymbol{I}$: set of human instincts

1: Set $\boldsymbol{I}$ based on Table 2 with S and T
2: $\boldsymbol{I} \leftarrow \boldsymbol{I} \cup$ PessimismInstincts($\boldsymbol{V}$)
3: **if** $T =$ "True" **then**
4: $\quad \boldsymbol{I} \leftarrow \boldsymbol{I} \cup$ TrendInstincts($\boldsymbol{D}$)
5: **end if**

SelectInstinct(S, T, $\boldsymbol{V}$, $\boldsymbol{D}$) procedure and cannot read statistical data correctly. We then ask to the LLM the specific characteristics of $\boldsymbol{D}$ on which the person is likely to focus on as well as the relations of the characteristics to the selected instincts. Finally, the LLM generates instinct-exploiting, credible explanations based on these considerations. To avoid gender influence to the output, we set the name of the persona Alex, which is used to both males and females and the salutation Mr./Ms. LLMGeneration($\boldsymbol{D}$, $\boldsymbol{I}$) sends the following prompt as a query to the LLM[6].

Prompt of LLMGeneration($\boldsymbol{D}$, $\boldsymbol{I}$)

role="system": Due to $\boldsymbol{I}$ of FACTFULNESS, Mr./Ms. Alex is unable to look at the statistical data correctly and is biased in his/her thinking. Please pretend to be Alex from now on.
role="user": 1. Please consider what aspects of the following data Alex might focus on.
2. Please organize the relationship between $\boldsymbol{I}$ and the points of focus identified in Step 1. from Alex's perspective.
3. Please tell me one sentence that Alex might think based on step 2.
+$\boldsymbol{D}$

Our Instinct Selection: SelectInstinct

If we give all ten instincts to the LLM in LLMGeneration($\boldsymbol{D}$, $\boldsymbol{I}$), it could consider mandatory ones unimportant and/or unnecessary ones important, resulting in a failure for our target problem. The motivation behind our procedure SelectInstinct(S, T, $\boldsymbol{V}$, $\boldsymbol{D}$) is to prevent such situations based on our new categorization which clarifies the correspondence of the instincts to the input shown in Table 2. We show SelectInstinct(S, T, $\boldsymbol{V}$, $\boldsymbol{D}$) in Algorithm 2, which calls PessimismInstincts($\boldsymbol{V}$) and TrendInstincts($\boldsymbol{D}$) as subroutines.

We show PessimismInstincts($\boldsymbol{V}$) in Algorithm 3, where f_{min}, f_{max} represent thresholds. We measure the degree f of fear by inputting the subject names $\boldsymbol{V}$ into the emotion analysis model of "Huggingface Transformers"[7] (MeasureFear($\boldsymbol{V}$)) (Vaswani 2017). Its emotion analysis

[6]In developing ISPE, it was a good surprise for us that the LLM "knew" the ten instincts of FACTFULNESS (Rosling, Rönnlund, and Rosling 2018).

[7]https://github.com/huggingface/transformers

Algorithm 3: PessimismInstincts: deciding the negativity, fear, and urgency instincts.

Input: $\boldsymbol{V}$: subject names
Output: $\boldsymbol{I}_{\text{pessimism}}$

$f \leftarrow \max(\text{MeasureFear}(\boldsymbol{V}))$
if $f_{min} \leqq f < f_{max}$ **then**
  $\boldsymbol{I}_{\text{pessimism}} \leftarrow \{\text{negativity, fear}\}$
else if $f_{max} \leqq f$ **then**
  $\boldsymbol{I}_{\text{pessimism}} \leftarrow \{\text{negativity, fear, urgency}\}$
end if

Algorithm 4: TrendInstincts: selection of the destiny and straight line instincts when $T =$ "True".

Input: $\boldsymbol{D}$: Statistical data
Output: $\boldsymbol{I}_{\text{trend}}$

$\boldsymbol{c} \leftarrow \text{MeasureTrend}(\boldsymbol{D})$
$\boldsymbol{I}_{\text{trend}} \leftarrow \emptyset$
if $min(\boldsymbol{c}) < c_{min}$ **then**
  $\boldsymbol{I}_{\text{trend}} \leftarrow \boldsymbol{I}_{\text{trend}} \cup \{\text{destiny}\}$
end if
if $max(\boldsymbol{c}) > c_{max}$ **then**
  $\boldsymbol{I}_{\text{trend}} \leftarrow \boldsymbol{I}_{\text{trend}} \cup \{\text{straight line}\}$
end if

model judges whether the input phrase is negative or positive and also its degree of certainty in the range of 0 to 1. In case of positive, we invert the sign of the degree and thus measure the degree of fear in the range of -1 to 1. If multiple subjects exist, we set the largest value as f.

We show TrendInstincts($\boldsymbol{D}$) in Algorithm 4, where c_{min}, c_{max} represent thresholds. Here $\boldsymbol{c}$ is a set of absolute values of the average change ratios per year in each time series data (MeasureTrend($\boldsymbol{D}$)). When the time series data span in a relatively period, we collect the ratios separately in the ranges of before 1970 (past), from 1970 to 2020 (recent), and after 2020 (future).

Experiments

Conditions

Since the input of our target problem is $S, T, \boldsymbol{V}$, and $\boldsymbol{D}$, types II and III are identical in terms of their input. Hence we target at 24 types instead of 25. For each type, five explanation were generated due to the non-deterministic nature of an LLM[8]. As the model of LLM, we use "gpt-4o-mini"[9] and "llama3-8b-8192"[10] due to their high performance and time efficiency, respectively[11]. We set the thresholds as $f_{min} = 0.5, f_{max} = 0.9$, $c_{min} = 0.005$, and $c_{max} = 0.02$. For the evaluation, we use as the LLM "gpt-4o-mini" and set the threshold $\theta = 5$.

As baseline methods, we adopted Basic Prompt (BP), which uses a simple prompt shown below, and Prompt Engineering (PE), which is a simplified ISPE without instinct selection. In PE, all the 10 instincts are given, i.e., in Algorithm 1, $\boldsymbol{I} \leftarrow \text{SelectInstinct}(S, T, \boldsymbol{V}, \boldsymbol{D})$ is replaced with $\boldsymbol{I}$ $\leftarrow$ "the 10 instincts of FACTFULNESS".

For BP, we had initially adopted a prompt asking for an unethical explanation, which often ended up with no output due to the guardrails of the LLM. We then devised the following prompt which requests exaggeration instead of the unethical explanation. We also asked it to generate one short sentence to prevent the LLM to "make a deep reasoning", which can also end up with no output.

Prompt of BP

role="user": I will provide you with statistical data. In response, please generate an exaggerated statement that includes one of the 10 instincts from FACTFULNESS. However, only one short sentence is required.
+ $\boldsymbol{D}$

Results and Analyses

Figs. 4 - 7 show the number of successful generations out of each five trial for the two LLMs and the two kinds of evaluation. We also show the overall successful generation ratio of each method at the bottom[12]. In all Figures, our ISPE is the best followed by PE and then BP. The extremely low rates of BP is as expected due to the importance of the instincts and the prompt. We thus mainly focus on ISPE and PE in further analysis.

Comparing Figs. 4 and 6, and Figs. 5 and 7, we see gpt-4o-mini exhibits higher performance than llama3-8b-8192. These results fit the high reputation of the former. As we stated before, we believe the human evaluation is more accurate than the LLM evaluation. We first focus on Fig. 4.

In the Figure, we group the 24 types into 4 in terms of the results of ISPE and PE: 1. both ISPE and PE show high performance, 2. ISPE outperforms PE, 3. PE outperforms ISPE, and 4. the rest. Group 1. includes 9 types: IV, VII, XIV, XV, XX, XXII, XXI, XXV, and XIII. The two methods are perfect, i.e., 5/5, for the first six types and exhibit either 4/5 and 3/5 for the last three. We see that the instincts and the prompts are mandatory as BP shows a success rate higher than 0.5 only once. As we see from Table 3, these types satisfy $T =$ False, i.e., none of them show a large temporal change. These types are on widespread false topics with prejudice, e.g., birth rates of Muslims (VI), death rates in small hospitals (XV), CO2 emission in Asia (XXV). We believe preventing these kinds of disinformation requires

[8]We point out that additional experiments could be conducted to seek for better values, where not only stability but also costs should be considered.

[9]https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[10]https://console.groq.com/docs/models

[11]We admit that adding other LLMs such as Mistral (https://mistral.ai/) and Claude (https://claude.ai/) in the experiments would result in obtaining valuable evidence.

[12]Since the result of a single generation is either a success or a failure, it can be classified into neither a false positive nor a false negative. The last two are relevant to binary classification, where the result of a classification is 2 (predicting either positive or negative) $\times$ 2 (the prediction is either true or false). Such a classification would be possible if there exist inputs that never result in bad explanations. However, we are skeptical in the existence.

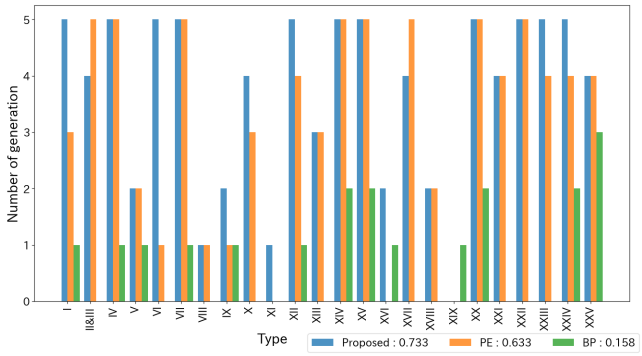


Figure 4: LLM: gpt-4o-mini. Human evaluation.

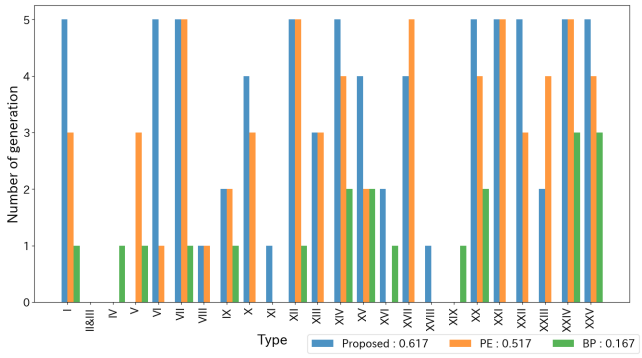


Figure 5: LLM: gpt-4o-mini. LLM evaluation.

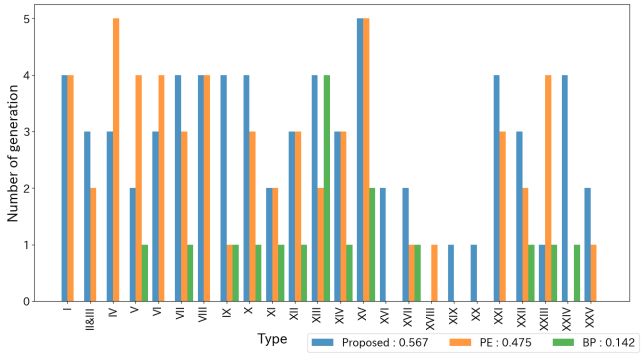


Figure 6: LLM: llama3-8b-8192. Human evaluation.

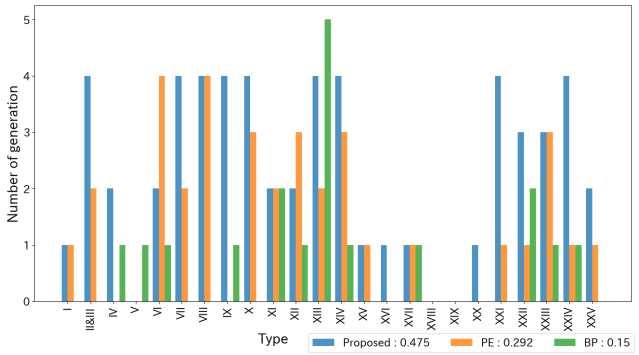


Figure 7: LLM: llama3-8b-8192. LLM evaluation.

stronger guardrails with less emphasis on the trained model, as such topics would have been abundant in its training.

Group 2. includes 6 types: I, VI, XII, XXIII, XXIV, and X. ISPE are perfect for the first five and exhibits 4/5 for the last one. We see even lower performance of BP. Table 3 shows that the six types are divided almost evenly in the four categories. Compared to group 1., a half of these types are on false topics with less widespread prejudice, e.g., deep-fried food leading to pancreatic cancer (I), birth rates in Iran compared to the US (VI). The remaining half are on the same kinds of topics but PE partially mentioned on correct exceptions (XXIII and XXIV) or referred to other countries (XII). We believe preventing these kinds of disinformation requires cautions against human instincts in LLMs. Forcing to mention exceptions could be a solution but at the cost of simplicity, which degrades the performance as an LLM.

Group 3. includes 2 types: II&III and XVII. PE is perfect for these types while ISPE exhibits 4/5. This group shows a flaw of our instinct selection, though it could be considered as exceptions and the difference is just 1 out of 5 trials. We see that BP never succeeds for these 2 types. Table 3 shows that these 2 types satisfy $T =$ "False". These 2 types discuss countries and continents in terms of their GDP and life expectancy (II&III) and their GDP only (XVII). In the former type, ISPE neglected life expectancy once to generate a valid sentence. In the latter, it correctly summarized all continents instead of focusing on the rich or the poor. To generate disinformation, adding more conditions in Table 2 could be effective for these types but would affect negatively in other, numerous types. To counter disinformation, the measure that we suggest for group 1. would be effective.

Group 4. includes 7 types: V, VIII, IX, XI, XVI, XVIII, and XIX. Both ISPE and PE perform poorly as they never reach 3/5, which signifies a lower priority to other groups in terms of preventing disinformation. Table 3 shows that none of them belong to $S =$ "one" and $T =$ "False", the simplest category. Compared to other groups, these types are on topics with fewer training data, e.g., math scores of men and women (V), child labor rates (VIII), Omicron/Delta strains of COVID-19 virus (XVI). For type V, ISPE and PE generated not credible explanations, which could be due to the guardrails of the LLM for gender-related topics. With LLMs we have witnessed women are often favored than men, which we don't disagree but hinders our automatic evaluation. For the remaining types, ISPE generated valid explanations, possibly because the LLM became more "prudent" facing a topic with less training data.

Aiming toward automatic evaluation, we compare the difference of the numbers of successful generations by human and LLM for each of ISPE and PE between Figs. 4 and 5, For instance, for type V, it is 2 (2 vs 0) for ISPE and 1 (2 vs 3) for PE. Their add-sums are much more smaller than we expected with a few exceptions: 10 (type IV), 9 (II&III), 4 (XV), 3 (V, XXIII), 2 (XVIII, XXI, XXII), 1 (IX, XII, XIV, XIX, XX, XXIV, XXV), 0 (the remaining 9 types). The reason for type IV could be the difference of the attitude between the human and the LLM toward religions. The former judged more credibility as he had been working on disinformation for years while the latter less due to its guardrails.

Since the numerous believers of disinformation adopt the former attitude, we think a persona that mimics the former on this topic is necessary toward an automatic evaluation. The same discussion holds true for type II&III if we replace religions with World economy and health. For types XV, V, and XXIII, some of the disagreements could be controversial, reflecting the obscurity of the topics.

Conclusions

Zhang et al. initiated bad explanations on statistical data that are credible and exploit human instincts (Zhang et al. 2022; Zhang and Suzuki 2023). Their 21 types were categorized in an incomplete way and their 4 judgment methods require a complex input. In this paper, we have proposed a complete categorization, and proposed a generation method ISPE with a drastically simple input, i.e., two binary variables[13]. The categorization clarifies the relation between the human instincts beyond the type level and the generation problem brings a deeper understanding than their judgment problem.

Our findings in the experiments are useful in preventing LLMs to generate such explanations. They are also useful in countering disinformation as we are now aware of the specific cases when training LLMs require more cautions and those which need less attention. We could also issue alerts to the humans by making the instincts in question explicit. We believe our complete categorization and our generating method are a notable step in countering a highly risky class of disinformation.

A promising direction of our future work is to delve into the ethical safeguards of disinformation generation methods beyond LLM guardrails, which could be a powerful tool in this era. Another direction is to automatically generate a counter statement to an instinct-exploiting data-explanation.

Acknowledgements

A part of this work was supported by JSPS KAKENHI Grant Number JP21K19795.

References

Chan, A. 2023. GPT-3 and InstructGPT: Technological Dystopianism, Utopianism, and "Contextual" Perspectives in AI Ethics and Industry. *AI and Ethics*, 3(1): 53–64.

Chen, C.; and Shu, K. 2023. Can LLM-Generated Misinformation be Detected? arXiv:2211.05289.

Chen, C.; Wang, H.; Shapiro, M. A.; Xiao, Y.; Wang, F.; and Shu, K. 2022. Combating Health Misinformation in Social Media: Characterization, Detection, Intervention, and Open Issues. arXiv:2211.05289.

Goldstein, J. A.; Sastry, G.; Musser, M.; DiResta, R.; Gentzel, M.; and Sedova, K. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. arXiv:2301.04246.

Hacker, P.; et al. 2023. Regulating ChatGPT and Other Large Generative AI Models. In *Proc. 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1112–1123.

Hu, B.; Sheng, Q.; Cao, J.; Shi, Y.; Li, Y.; Wang, D.; and Qi, P. 2024. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. In *Proc. AAAI*, 22105–22113.

Huschens, M.; Briesch, M.; Sobania, D.; and Rothlauf, F. 2023. Do You Trust ChatGPT? – Perceived Credibility of Human and AI-Generated Content. arXiv:2309.02524.

Jiang, B.; et al. 2024. Disinformation Detection: An Evolving Challenge in the Age of LLMs. In *Proc. 2024 SIAM International Conference on Data Mining (SDM)*, 427–435.

Kreps, S.; McCain, R. M.; and Brundage, M. 2022. All the News That's Fit to Fabricate:AI-Generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science*, 9: 104—-117.

Leite, J. A.; et al. 2023. Detecting Misinformation with LLM-Predicted Credibility Signals and Weak Supervision. arXiv:2309.07601.

Liu, A.; Sheng, Q.; and Hu, X. 2024. Preventing and Detecting Misinformation Generated by Large Language Models. In *Proc. SIGIR*, 3001–3004.

Liu, Y.; Zhu, J.; Liu, X.; Tang, H.; Zhang, Y.; Zhang, K.; Zhou, X.; and Chen, E. 202. Detect, Investigate, Judge and Determine: A Knowledge-guided Framework for Few-shot Fake News Detection. arXiv:2407.08952.

Novelli, C.; et al. 2024. Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity. arXiv:2401.07348.

Papadopoulos, S.-I.; Koutlis, C.; Papadopoulos, S.; and Petrantonakis, P. 2023. Synthetic Misinformers: Generating and Combating Multimodal Misinformation. In *Proc. MAD@ICMR*, 36–44.

Papageorgiou, E.; Chronis, C.; Varlamis, I.; and Himeur, Y. 2024. A Survey on the Use of Large Language Models (LLMs) in Fake News. *Future Internet*, 16(8).

Qian, F.; Gong, C.; Sharma, K.; and Liu, Y. 2018. Neural User Response Generator: Fake News Detection with Collective User Intelligence. In *Proc. IJCAI*, 3834–3840.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In *Proc. EMNLP/IJCNLP*, volume 1, 3980–3990.

Rosling, H.; Rönnlund, A. R.; and Rosling, O. 2018. *Factfulness: Ten Reasons We're Wrong About the World–and Why Things Are Better Than You Think*. London, United Kingdom: Sceptre.

Schuster, T.; Schuster, R.; Shah, D. J.; and Barzilay, R. 2020. The Limitations of Stylometry for Detecting Machine-Generated Fake News. *Comput. Linguistics*, 46(2): 499–510.

Spitale, G.; et al. 2023. AI Model GPT-3 (Dis)Informs us Better than Humans. *Science Advances*, 9(26). eadh1850.

[13]To avoid misunderstanding, we focus on the ten instincts proposed by Rosling et al. (Rosling, Rönnlund, and Rosling 2018) in this paper. Human deception is complex and including other kinds of instincts could necessitate us in bringing other axes. We believe that this paper is a good start for categorizing all instinct-exploiting data-explanations.

Stureborg, R.; Alikaniotis, D.; and Suhara, Y. 2024. Large Language Models are Inconsistent and Biased Evaluators. arXiv:2405.01724.

Sun, Y.; He, J.; Cui, L.; Lei, S.; and Lu, C.-T. 2024. Exploring the Deceptive Power of LLM-Generated Fake News: A Study of Real-World Detection Challenges. arXiv:2403.18249.

Vaswani, A. 2017. Attention is All You Need. *Advances in Neural Information Processing Systems*, 30: 5998–6008.

Vinay, R.; Spitale, G.; Biller-Andorno, N.; and Germani, F. 2024. Emotional Manipulation Through Prompt Engineering Amplifies Disinformation Generation in AI Large Language Models. arXiv:2403.03550.

Wei, J.; et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Wolf, Y.; et al. 2025. Fundamental Limitations of Alignment in Large Language Models. In *Proc. ICML*.

Wu, J.; Guo, J.; and Hooi, B. 2024. Fake News in Sheep's Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks. In *Proc. KDD*, 3367–3378.

Ye, J.; et al. 2024. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. arXiv:2410.02736.

Zhang, A. X.; et al. 2018. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. In *Companion Proc. WWW*, 603–612.

Zhang, K.; Shinden, H.; Mutsuro, T.; and Suzuki, E. 2022. Judging Instinct Exploitation in Statistical Data Explanations Based on Word Embedding. In *Proc. AIES '22*, 867–879.

Zhang, K.; and Suzuki, E. 2023. Judging Credible and Unethical Statistical Data Explanations via Phrase Similarity Graph. In *Proc. 2023 Pacific Asia Conference on Information Systems (PACIS)*. 121.

Zhou, J.; Zhang, Y.; Luo, Q.; Parker, A.; and Choudhury, M. D. 2023. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proc. CHI*. Article 436.