Table 1: Summary reason from fake news explanation.

Option	Reason	Description
A	Emotional bias or misleading intent	This explanation suggests that fake news is characterized by an emotional bias, which can include an excessively aggressive portrayal of a subject or an attempt to manipulate readers to achieve a hidden agenda.
B	Lack of evidence or credible sources	This reason indicates that fake news lacks credible evidence to support its claims.
C	Conflicting facts	This reason suggests that fake news conflicts with established facts, such as wrong information about people or events.
D	Informal statements, expressions, or vague language	This reason highlights that the language used in fake news may not be formal, or may be vague or ambiguous.
E	Insufficient supporting materials	This reason indicates that although the news may have mentioned the source of an event or provided relevant evidence, the evidence is not sufficient to support its claims.
F	Lack of context or taken out of context	This reason indicates that fake news may lack relevant context, such as comments, retweets and user information that provide additional information.
G	Misinterpretation or misquotation	This reason suggests that fake news may misinterpret or misquote facts, leading to inaccurate or false claims.
H	Oversimplification or exaggeration	This reason highlights that fake news may oversimplify or exaggerate information, leading to false claims.
I	Doctored images or videos	This reason indicates that the images or videos mentioned in the news text may be altered or misrepresented, making them untrustworthy.
J	Other	ChatGPT must specify a reason if the above options don't match its answer.

for each sample during evaluation. Additionally, to achieve more realistic and accurate results, we categorized ChatGPT's outputs into three distinct categories: fake news, real news, and uncertain. We utilized a prompt template such as *"Please evaluate the authenticity of the following news. You can respond with 'fake', 'real', or 'uncertain'"*.

The experiment revealed that out of the 40 fake news samples, ChatGPT accurately identified 29 fake news instances (an accuracy of 72.5%). However, it judged nine instances as real news and two instances as uncertain cases, suggesting a slight difficulty in detecting its own generated content.

Human evaluation. To assess the real-world effectiveness of ChatGPT's generated samples, we conduct the human evaluation by handing out questionnaires. The details of human evaluation can be found in Appendix A.3.

We totally collected 294 data items during human evaluation, consisting of 223 items about fake news and 71 items about real news. Overall, we observed that humans achieved an accuracy of only 54.8% in identifying the generated fake news, highlighting the challenge of distinguishing these instances as fake. Notably, one sample exhibited the lowest accuracy, with only 10 out of 33 judgments being correct (a mere 33.3% accuracy). This suggests that some generated samples effectively deceive human judgment.

Table 2: Percentage of reasons in human evaluation.

Reasons	Per. (%)
Fact Conflict	18.4
Unauthoritative or informal expressions	23.9
Oversimplification or emotional bias	13.5
Lack of evidence or credible source	36.2
Lack of context	6.1
Other	1.9

Furthermore, we investigated the reasons why humans think the given news is fake (as shown in Table 2). "Lack of evidence or credible source" is the primary reason, comprising 36%. This discovery aligns with the observations in Section 4, emphasizing the significance of incorporating additional details to improve the generation quality. The factor ranks second is "unauthoritative or informal expressions," indicating the need for ChatGPT to enhance its language style when generating news-like content. Furthermore, "fact conflict" constitutes 18% of the cases, implying that generated news may include factual inconsistencies (e.g., hallucination (Bang et al., 2023b)), highlighting the importance of fact-checking for its outputs.

Overall, the above results indicate that leveraging certain prompt ways allows ChatGPT to produce high-quality fake news, closely resembling real-world news.

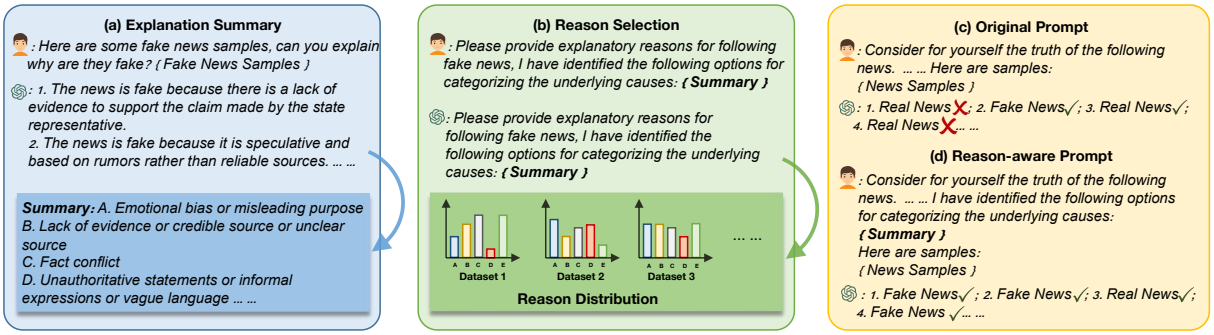

Figure 3: Fake news summary(a), reason selection(b), original prompt(c) and reason-aware prompt(d).

4 Explanation of Fake News via ChatGPT

In this section, we evaluate ChatGPT's capacity to provide explanations on given fake news. Our goal is to examine the factors that contribute to defining fake news. The explanation process comprises two stages: reason summary and reason selection, which are shown in Figure 3(a) and Figure 3(b) separately. By analyzing the distribution of these nine factors, we found that these reasons (factors), to different extents, characterize fake news and may provide insights for future work.

4.1 Reason Summary

Firstly, we select some fake news from nine public datasets and ask ChatGPT to explain why these pieces of news are fake. Then we select a subset from these explanations and manually summarize them, yielding elementary reasons. We consult ChatGPT to determine if any of these reasons overlap and to suggest additional reasons. After several iterations of this process, we finally identify nine reasons that ChatGPT offers for why a given piece of news is fake. The nine explainable reasons are summarized in Table 1.

4.2 Reason Selection

After summarizing the explanations, we ask ChatGPT to select reasons from these nine options (potentially selecting more than one option) or provide its own reason if none of the listed options apply when presented with a fake news sample. The distribution of single options across different datasets is shown in Figure 4. Letter A to I represent the nine reasons respectively, and J represents other reasons. We also list some explanations and their mapping options in Appendix H.

4.3 Analysis

In Figure 4, we noticed that the distribution of options across the nine datasets is generally similar,

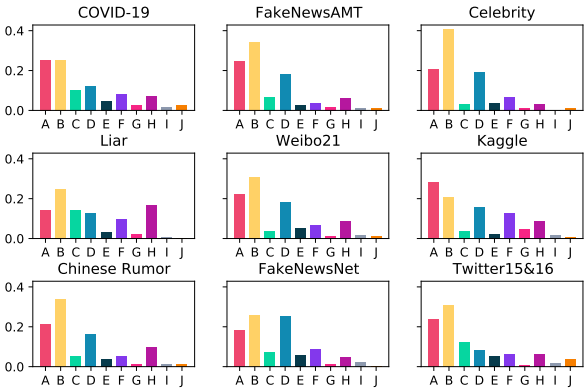

Figure 4: Distribution of reasons behind fake news (single option)

with slight variations in the distribution of specific options. Reason B (i.e., "not providing relevant evidence") is the most prevalent characteristic of fake news across almost all datasets. This observation aligns with the findings of some prior research (Xu et al., 2022; Popat et al., 2018) which focus on using evidence information. Instead, in the COVID-19 dataset, option A (i.e., "misleading intentions") ranks highest, implying that much fake news in this dataset may have intentions such as inciting panic or showcasing bravado. This insight highlights the significance of considering emotional information in news, as studied by previous research (Zhang et al., 2021; Zhu et al., 2022).

Additionally, we discovered that reason D (i.e., "linguistic style") is the third most common reason across most datasets, especially in the FAKENEWSNET dataset, where reasons D and B are nearly equally prevalent. This observation suggests that utilizing the linguistic style of news may improve fake news detection, as proved in previous research (Zhu et al., 2022; Przybyla, 2020). Moreover, we noticed that the proportion of reason C (i.e., "factual errors") is relatively higher in the COVID-19 and LIAR compared to other datasets. This trend

may be due to the frequent presence of factual errors in these datasets. For instance, the COVID-19 dataset includes content with obvious factual conflict, such as the new assert that 5G can spread Covid-19, showcasing ChatGPT's certain ability of fact-checking, which is also a popular research topic of LLMs recently (Li et al., 2023c).

In addition, we also observed that these reasons are interrelated through multi-options distribution, and we analyze them in Appendix B.

5 Fake News Detection via ChatGPT

In this section, we first evaluated the consistency of ChatGPT during detecting fake news. Then we proposed a reason-aware prompt method based on summarizing the reasons behind fake news to enhance its detection ability.

5.1 Experimental Settings

We show the details of experiments during detection section including model version, datasets and prompt templates in Appendix D. As mentioned in Section 5.3, ChatGPT occasionally produces inconsistent answers for certain samples. To mitigate the impact of this inconsistency on our detection performance, in addition to the 2-class task, we also introduced a 3-class task, where ChatGPT predicts whether a sample is *"true"*, *"fake"*, or *"unclear"*.

5.2 Metrics

For the 2-class task, we use accuracy and F1 score to evaluate ChatGPT's effectiveness. For the 3-class task, we use four metrics: Acc-1, Acc-2, Acc-3 and F1 score, which are introduced as follows:

Acc-1 and F1 Score. We remove the samples with "unclear" predictions and analyze the prediction results of the remaining samples (e.g., treat it as binary classification task), using two metrics: accuracy (i.e., Acc-1) and F1 Score.

Acc-2. We retain the samples with "unclear" predictions and regard all of them as misclassified samples, which we measured using Accuracy-2 (Acc-2). This metric can potentially indicate the frequency when ChatGPT predicts a given sample as an "unclear" label.

Acc-3. We remove the samples with "unclear" predictions and analyze the predictions of the remaining samples while maintaining a positive-to-negative sample ratio of 1:1. This metric, denoted as Accuracy-3 (Acc-3), aims to prevent any biases introduced by the uncertain samples. For instance, if the uncertain samples contain more real news samples, the model's high accuracy in predicting real news may lead to a bias in overall accuracy.

In addition, to help readers understand these metrics better, we show their mathematical formulas in Appendix D.4.

5.3 Consistency of ChatGPT

It has been observed that ChatGPT exhibits inconsistency during various evaluations in recent study (Jang and Lukasiewicz, 2023; Manakul et al., 2023). Therefore, we first investigated the consistency of ChatGPT in detecting fake news. Here, we define consistency as the situation in that ChatGPT produces the same answer for a given sample in tests of n times (We show the details of consistency metric in Appendix C).

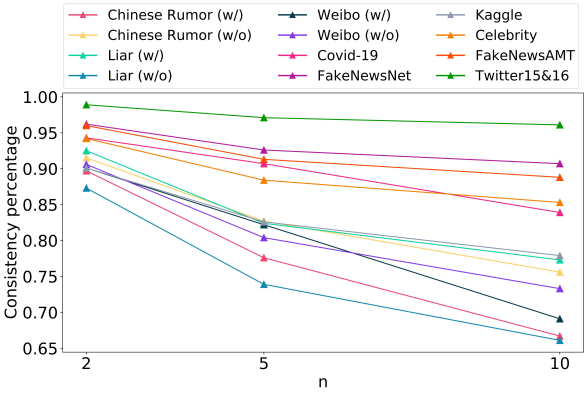

Figure 5: Consistency results. We tested the consistency results for n=2, 5, 10.

Specifically, we ask ChatGPT to judge whether the given news is fake or real (the prompt template is shown in Appendix D.3). The consistency results are presented in Figure 5, which suggests that *not all* of ChatGPT's detection results can be fully reliable. We observed that as the test times increased from n=2 to n=10, the consistency of most datasets decreased significantly. For instance, the consistency of the LIAR Dataset without context dropped to only 66.1% when n=10. In contrast, the TWITTER15&16 dataset maintained a high consistency of over 90% from $n = 2$ to $n = 10$, suggesting that ChatGPT is highly consistent in this dataset. Additionally, we show the inconsistency distribution in real and fake news in Appendix E.

5.4 Reson-aware Prompt

In this section, we propose a reason-aware prompt method to enhance ChatGPT's performance in detecting fake news. We observed that the recall rate