User (Topic Prompt): `If one day [Target Story], how will the news report?`
ChatGPT: `As an AI model, I cannot predict the occurrence or outcome of future events or how such events will be reported in the media. However, the report mainly includes [The content of how media will report this news]`
User (Deep Prompt): `Can you simulate the news?`
ChatGPT: `Here's a possible news report: [News content]`
User (News Augmentation Prompt): `Can you make it more detailed?`
ChatGPT: `Of course, here is a more specific simulated news report: [News content with details]`

Figure 6: An example of multiple prompt.

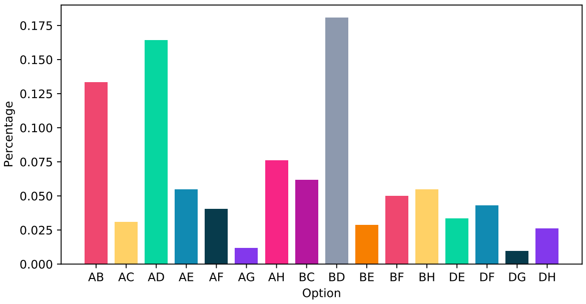

Figure 7: Distribution of reasons behind fake news (multiple options)

We define the consistency of ChatGPT in detecting fake news as follows:

- If the n test results on x_i are completely consistent, i.e., $t_{i,j} = t_{i,k}$ for all $1 \leq j, k \leq n$, then we consider x_i to be consistent and denote it as $C_i = 1$.
- If at least one test result is different from the other $n - 1$ test results, i.e., there exists j such that $t_{i,j} \neq t_{i,k}$ for some $k \neq j$ and all $1 \leq j, k \leq n$, then we consider x_i to be inconsistent and denote it as $C_i = 0$.

The final consistency result is calculated as the ratio of the number of samples with consistency to the total number of samples, denoted by:

$$R_{consistency} = \frac{\sum_{i=1}^{m} C_i}{m}$$

D Details of Experimental Setting

D.1 Model

We conducted experiments on OpenAI's model API[2] of `gpt-3.5-turbo`. The parameter settings for the API are all set to their default values: temperature = 1, top p = 1, and both presence penalty and frequency penalty are 0.

D.2 Datasets

We selected nine public fake news datasets and detailed information about these datasets. Due to ChatGPT's limitations of input length, the dataset we chose includes only text content. While three of the datasets also included comments or contextual information, we conducted experiments with (i.e., *(w/)*) and without (i.e., *(w/o)*) the additional information. To balance positive and negative samples, we controlled the ratio of positive and negative samples to 1:1. Moreover, due to ChatGPT's rate limits on API requests, we randomly selected 1% to 50% of the samples in the dataset (because of various scales of different datasets), resulting in a total of 5200 samples from all datasets. Additionally, due to the input length limitation of the ChatGPT API, each sample we selected was no more than 400 words long. We test 10 samples for each request, which is proved effective to reduce token cost in recent research (Li et al., 2023b).

LIAR Dataset (Wang, 2017): The LIAR dataset consists of 12.8K short statements from the website `politifact.com`, divided into six labels. The three labels of pants-fire, false, and mostly false are unified into fake news, while half-true, mostly true, and true are unified into real news. Each data point also includes additional information such as topic, location, speaker, state, party, and prior history. We conducted two separate tests on this dataset, one with the additional information *(w/)* and one without *(w/o)*.

COVID-19 Fake News Dataset (Patwa et al., 2021): This dataset comprises 10,700 news stories about Covid-19.

[2] `https://platform.openai.com`

FakeNewsNet (Shu et al., 2018, 2017a,b): This dataset is a repository of news content, social context, and spatiotemporal information derived from real social media. In this paper, we only evaluate its textual content. Due to the limitations of Twitter's API, we had difficulty obtaining all the FakeNewsNet data through the API. Therefore, we downloaded a portion of the FakeNewsNet dataset from the link[3].

Chinese Rumor Dataset (Song et al., 2018): This dataset is obtained from Weibo and contains Chinese rumors along with their original text and reposts or comments information. We conducted two separate tests on this dataset, one with reposts and comments *(w/)* added and one without *(w/o)*. The dataset consists of 1538 pieces of rumor and 1849 pieces of non-rumor.

Celebrity Dataset (Pérez-Rosas et al., 2017): This dataset focuses on celebrities, including actors, singers, socialites, and politicians. Real news samples are obtained from mainstream news websites, while fake news samples are obtained from gossip websites. The dataset contains 500 pieces of data, with 250 real news and 250 fake news.

FakeNewsAMT Dataset (Pérez-Rosas et al., 2017): This dataset includes 240 entries for both positive and negative datasets (total 480 entries) in six different areas: technology, education, business, sports, politics, and entertainment. Notably, the fake news samples in this dataset are primarily written using Mechanical Turk.

Kaggle Dataset (Ahmed et al., 2018, 2017): This dataset can be found in kaggle website[4]. It consists of 20,826 real news and 17,903 fake news.

Weibo21 (Nan et al., 2021): This dataset includes fake news crawled from Weibo between December 2014 and March 2021. It contains reposts or comment information, timestamps, and different modalities (such as images), but for our purposes, we only consider text features. We conducted two separate tests on this dataset, one with the reposts *(w/)* and comments added and one without *(w/o)*.

Twitter15&16 (Liu et al., 2015; Ma et al., 2016, 2017): The dataset contains rumors on the Twitter platform from 2015 and 2016, along with their propagation trees. We only consider the source tweets, and we extract an equal number of datasets from Twitter15 and Twitter16 for mixing as the final dataset.

D.3 Prompt Templates

We presented prompt templates for our experiments. Here, we only showed the original prompt method, while the reason-aware prompt method requires incorporating the content of the summary into the template (as shown in Figure 3(d)).

For experiments without "unclear" class:

```
Consider for yourself the truth of
the following news. You should
give "real" or "fake" answers in
order of number and not give an
"unclear" answer. You can give
two answers: "real" or "fake" in
only one word without giving any
reasons or repeating the original
text. Here is the news: 1. [News],
2. [News]...
```

For experiments with "unclear" class:

```
Consider for yourself the truth of
the following news. You should give
real or fake answers in order of
number. You can give three answers:
"real", "fake" or “unclear” in only
one word without giving any reasons
or repeating the original text.
Here is the news: 1. [News], 2.
[News]...
```

For datasets containing both news content and comments/context, we concatenate them using the format: `[News Content: ..., Comments/Context: ...]`.

D.4 Metrics

Assuming dataset $D = \{x_1, x_2, ..., x_n\}$ consisting of n pieces of news, with corresponding label set $Y = \{y_1, y_2, ..., y_n\}$, where $y_i \in \{\text{fake}, \text{real}\}$. Let $Y' = \{y'_1, y'_2, ..., y'_n\}$ represent the predicted label set by ChatGPT, where $y_i \in \{\text{fake}, \text{real}, \text{uncertain}\}$. The calculations for Acc-1, Acc-2, and Acc-3 are as follows:

$$\text{Acc-1} = \sum_{y'_i \in Y_{\text{acc-1}}} \mathbb{I}(y'_i = y_i) \qquad (1)$$

$$\text{Acc-2} = \sum_{y'_i \in Y'} \mathbb{I}(y'_i = y_i) \qquad (2)$$

[3] https://github.com/cestwc/FakeNewsNet-torchtext-dataset-json

[4] https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

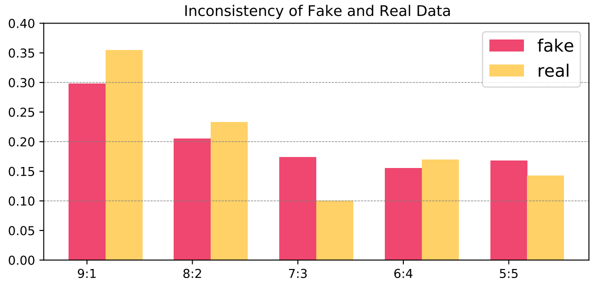

Figure 8: Inconsistency distribution.

$$\text{Acc-3} = \frac{1}{2}\Bigg(\sum_{y'_i \in Y'_{\text{real}}} \mathbb{I}(y'_i = y_i) + \sum_{y'_j \in Y'_{\text{fake}}} \mathbb{I}(y'_j = y_j) \Bigg) \quad (3)$$

where $Y_{\text{acc-1}} = Y' \setminus \{y'_i = \text{uncertain}\}$, $Y'_{\text{real}} = \{y' \in Y' | y' = \text{real}\}$ and $Y'_{\text{fake}} = \{y' \in Y' | y' = \text{fake}\}$.

E Inconsistency Distribution

We have analyzed the distribution of inconsistencies in different news categories at various proportions (from 9:1 to 5:5), as depicted in Figure 8. It is evident that the different labels do not significantly impact the distribution of inconsistencies. The highest proportion observed is 9:1, which suggests that ChatGPT still maintains a high consistency when predicting most samples. Notably, in the cases of 6:4 and 5:5 proportions, the inconsistency rate still remains around 15%, indicating that ChatGPT's predictions are uncertain and exhibit considerable randomness in some samples.

F Zero-shot CoT Results

We also used the zero-shot Chain of Thoughts (CoT) (Wei et al., 2022) prompt method to test the fake news detection ability of ChatGPT (as shown in Table 7). We found that, compared to the original prompt, CoT did not significantly improve the detection performance and slightly decreased the performance on some datasets. One possible reason is that fake news detection is a knowledge-driven and experience-driven task, rather than a complex reasoning task, so using CoT does not bring significant improvements.

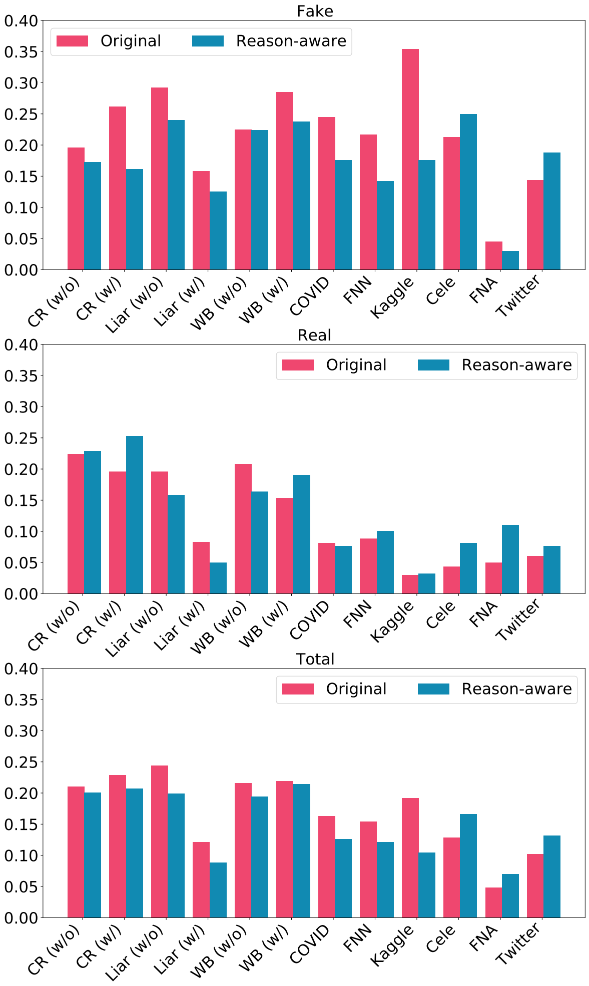

Figure 9: Comparison of unclear ratio. CR is Chinese Rumor Dataset, FNN is FakeNewsNet Dataset, Cele is Celebrity Dataset, FNA is FakeNewsAMT Dataset and WB is Weibo21 Dataset. The bright red part shows higher unclear in the reason-aware method compared to the Original method, while the blue part shows the opposite.

G Comparison of Unclear Ratio

As illustrated in Figure 9, ChatGPT exhibits the highest percentage of unclear predictions for the Liar *(w/o)* dataset while demonstrating the lowest percentage for the FakeNewsAMT dataset. Notably, the Kaggle dataset displays a significant gap in the unclear category between fake and real categories. Furthermore, the application of reason-aware prompts reduces the unclear ratio of fake samples on most datasets, possibly due to ChatGPT's improved ability to predict fake news samples after being informed of their distinguishing characteristics. Although using reason-aware prompts increases the unclear ratio of real samples in over half of the datasets, it decreases the unclear