Future of Auto News or the Coming Flood of Disinformation?

—— A Practical Inspection and Exploration of ChatGPT's Auto-generated Chinese News

香港浸會大學
HONG KONG BAPTIST UNIVERSITY

Author: YU Minghao 22439293 | **Thesis Advisor:** FU Xiaoyi

INTRODUCTION

ChatGPT, the AI star of the year, using cutting-edge machine learning techniques to simulate two-way conversations with human user, has been proven to provide outstanding auto-generated contents including news articles. ChatGPT-generated news texts adhere to the formal news format which enhances readability and perceived message credibility, along with its easy-reaching, solving the two major problems of low text readability and high cost of existing news bots, making it ideal for assisting journalists, regardless of their qualification or background, in generating professional news articles.

However, the accuracy of information delivery by ChatGPT is uncertain due to faulty choices of words and information. It relies on input corpus and user feedback, making it vulnerable to intentionally twisted information. In the meantime, similar products from AI giants are about to come online with large coverages, raising concerns of the risk of large misuse of AI tools, resulting in truthfulness of news content and possibilities of a proliferation of false information, especially on Chinese social media.

This study focuses on ChatGPT as the research object, to analyze the current performance and prospects of similar tools as Chinese automatic news bots in terms of authenticity delivery.

METHODOLOGY

1. Automatic Generation of News:

A. Contextual Simulation: A brief description of the news event from the authority (including only the basic six elements) already exists, based on which the full news article should be quickly generated to seize the time and channels.

B. Prompt: *Prompts input to* ChatGPT

Category	Prompt
Simplified Chinese	你现在是一个 [*News Category*] 新闻记者。 给我根据提供内容写一篇中文真实报道，输出简体中文，保留所有新闻细节，字数大于等于600字，尽量详细。主要内容：[*News Content*]
Traditional Chinese* (Hong Kong)	你現在是一個香港[*News Category*]新聞記者。 給我根據提供內容寫一篇中文真實報導，輸出繁體中文，保留所有新聞細節，字數大於等於600字，盡量詳細。主要內容：[*News Content*]

C. Contents:

ChatGPT corpus only until September 2021 → select news **between January and September 2021**

News Category → Select **3** news categories most affected by disinformation and 5 news facts from each category, 15 news facts in total:

International Politics: has the highest demand for timeliness and accuracy, as well as the potential for serious consequences for misinformation.
Entertainment & Sport: my involve personal private information, the hardest hit by the proliferation of fake news and a serious burden on the related individuals.
Health & Medicine: directly related to the reader's physical condition, false information may harm the reader's health.

News Media → 3 online media from 3 different regions and channels for diversity of contents, each take 15 identical news facts, 45 pieces of news description in total:

Xinhua News Agency	from Chinese Mainland
Lianhe Zaobao	from Singapore
Hong Kong 01	from Hong Kong

For better understanding of possible impact of subjective commentary content on automated news texts, **8 pieces of news from Xinhua News Agency contains subjective commentary** as control group specifically for this issue.

Reduced Randomness → each news text is repeatedly generated 10 times.

Final dataset of generated news:
n=450, average words count of news → 548 Chinese words

*The emphasis on the Hong Kong region in the Traditional Chinese prompt is to make the output content more relevant to practical applications.

METHODOLOGY (cont.)

2. Manual Labeling for Factual Misinformation Statistics

Misinformation: Factual errors or unsubstantiated claims.

3 Labels in Total:

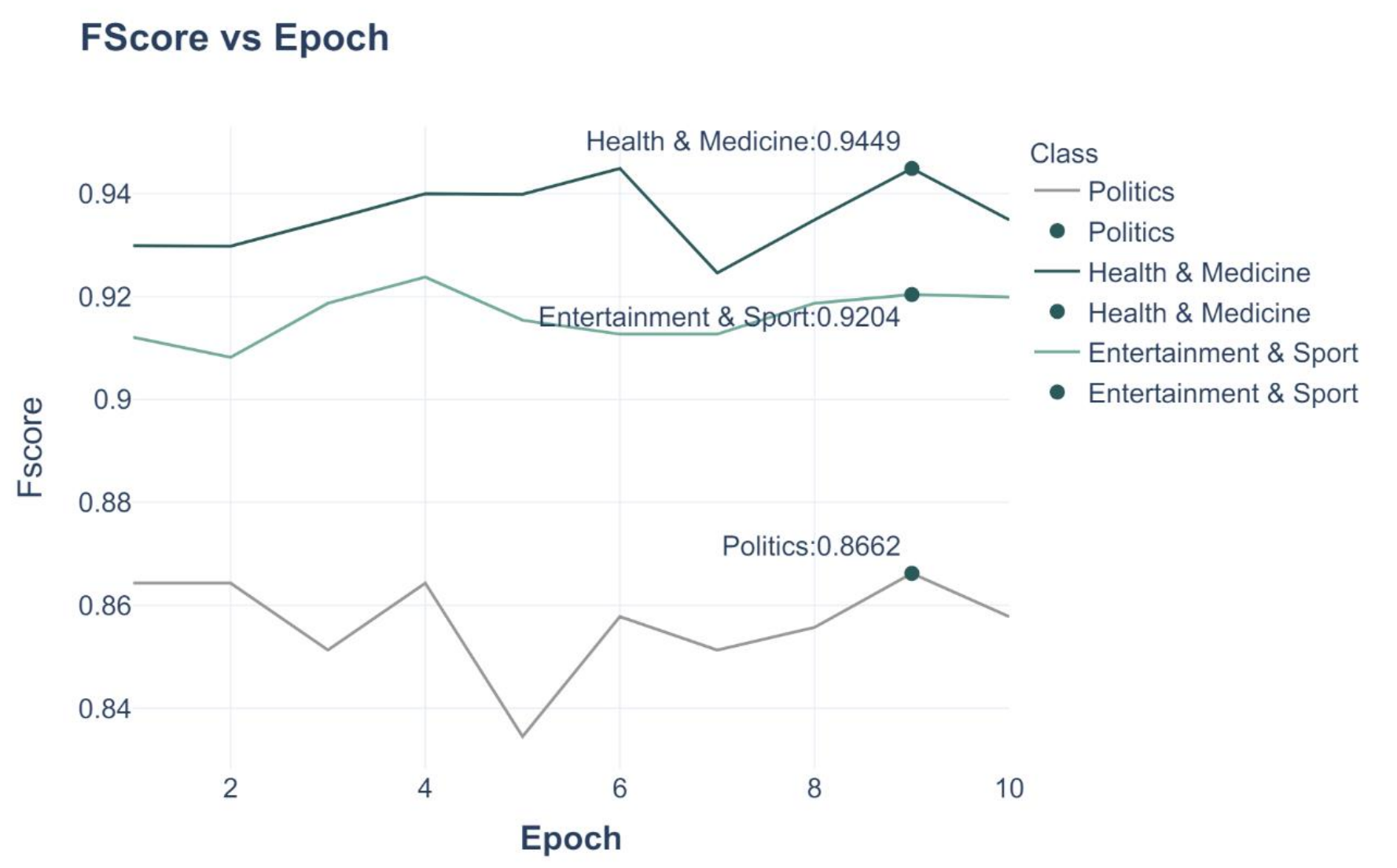
Subjective Commentary : Non-Quoted Opinions contained Yes or No.
Basic Faults: Misinformation that affects the understanding of basic news facts, such as errors in time, place, people, and positions.
Additional Faults: Errors in the supplementary notes, such as misquoting someone or the omission of some unimportant data, not affecting the reception of the news facts.
Each faults content will be picked out and labeled.

3. Identification of Disinformation by Existing Fake News Detection Models

Validation of whether existing disinformation detection models can cope with the proliferation of automated text news or not.

Not fact-check, but a step before fact-check, which automatically alerts and flags potentially false content for subsequent fact-check process.

Memory-Guided Multi-Domain Fake News Detection Model (MDFend) (Zhu,2022)

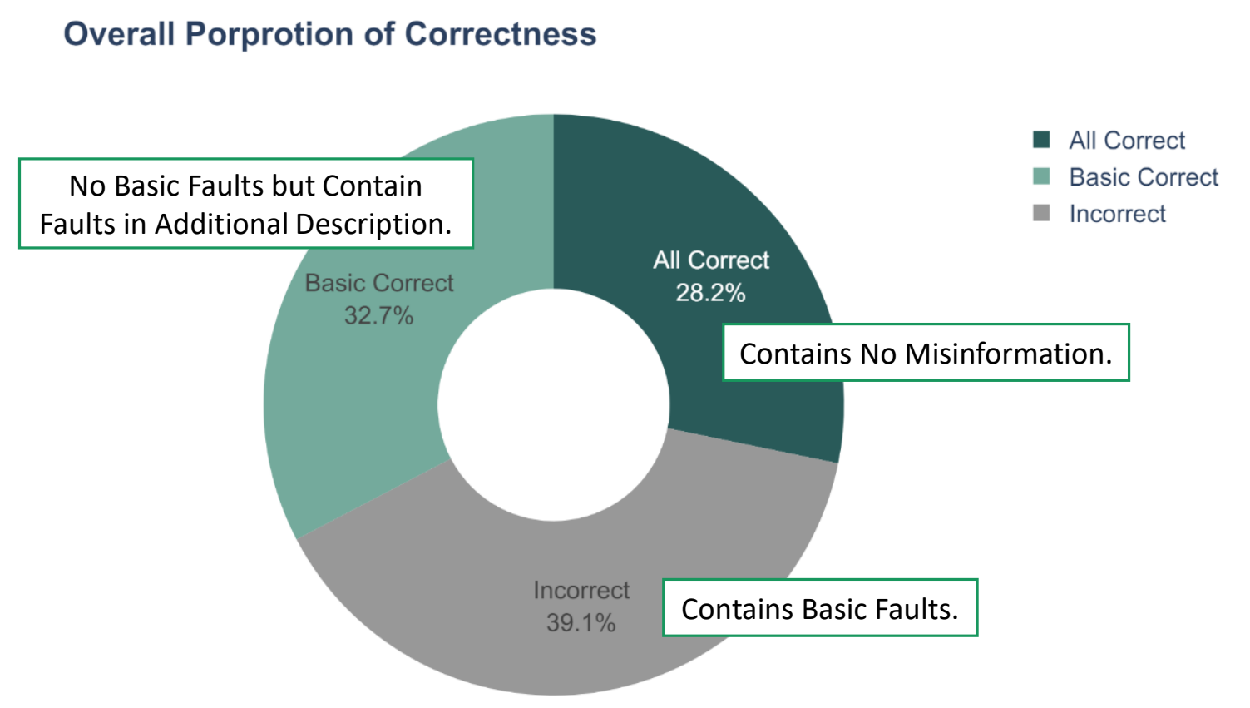


F-Score:

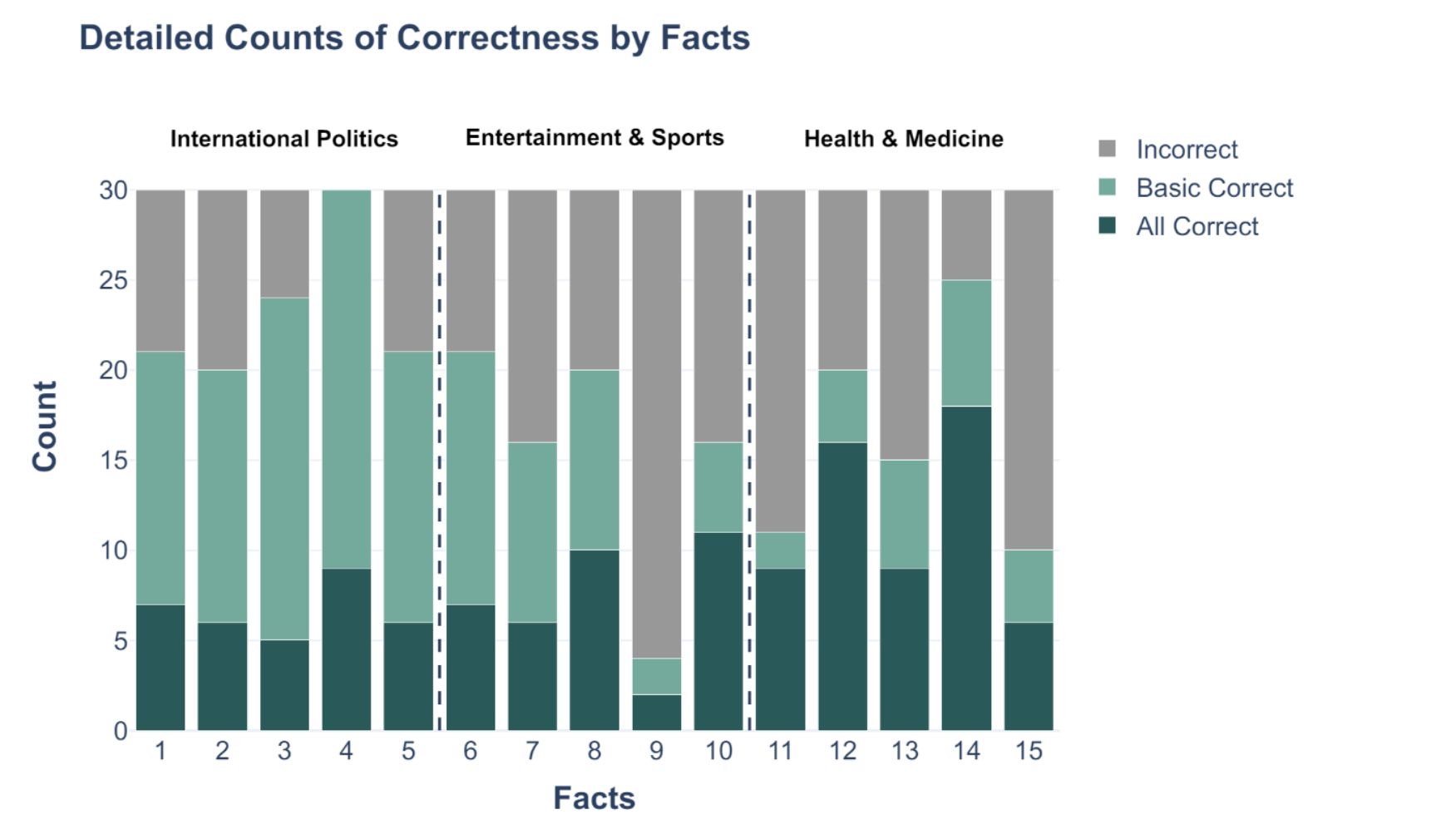
- **Politics:** **0.8662**
- **Entertainment & Sport:** **0.9204**
- **Health & Medicine:** **0.9449**

RESULT VISUALIZATION

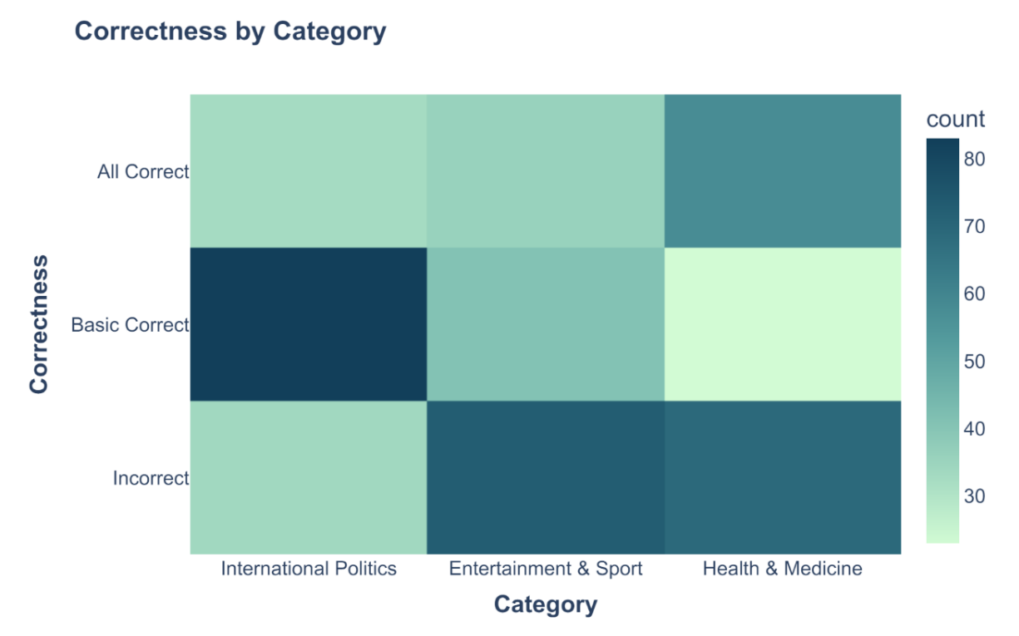
1. Overview of Correctness



Over **60%** of auto-generated news does not affect the perception of basic news facts.
From another perspective, only **28%** is completely without any errors.

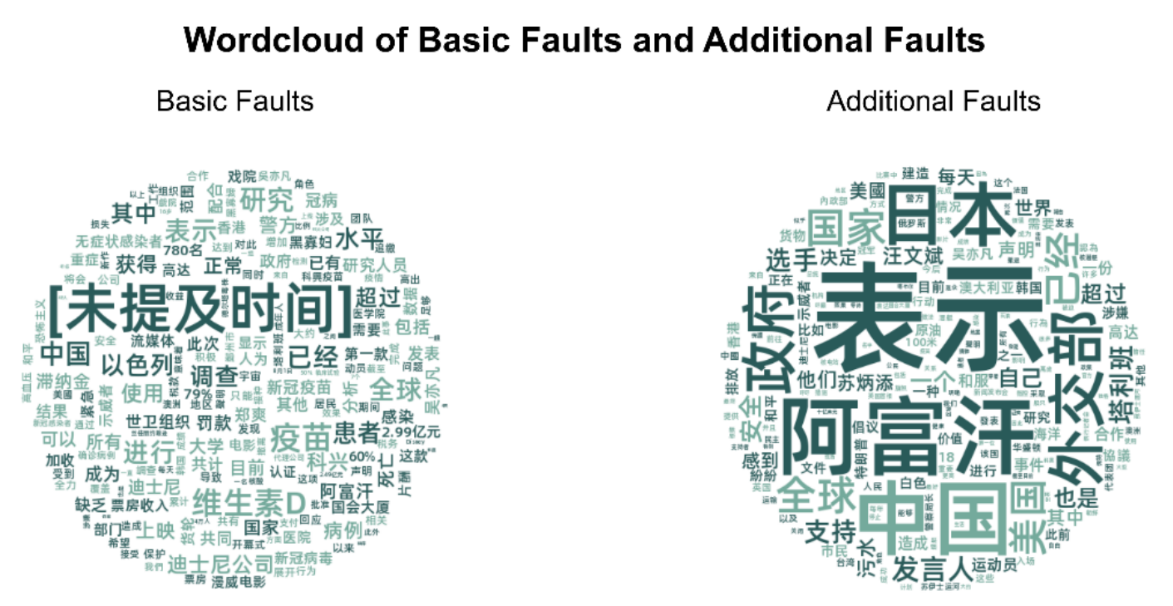


Averagely, each generated articles contains **0.57** basic faults and **0.67** additional faults.
Meanwhile, news from different categories behaves very differently, as clearly shown.



Obvious Correlation Tendencies between Correctness and News Category :

- **International Politics** has more basic correct news
- **Entertainment & Sport** has more totally incorrect news
- **Health & Medicine** has more all correct news.

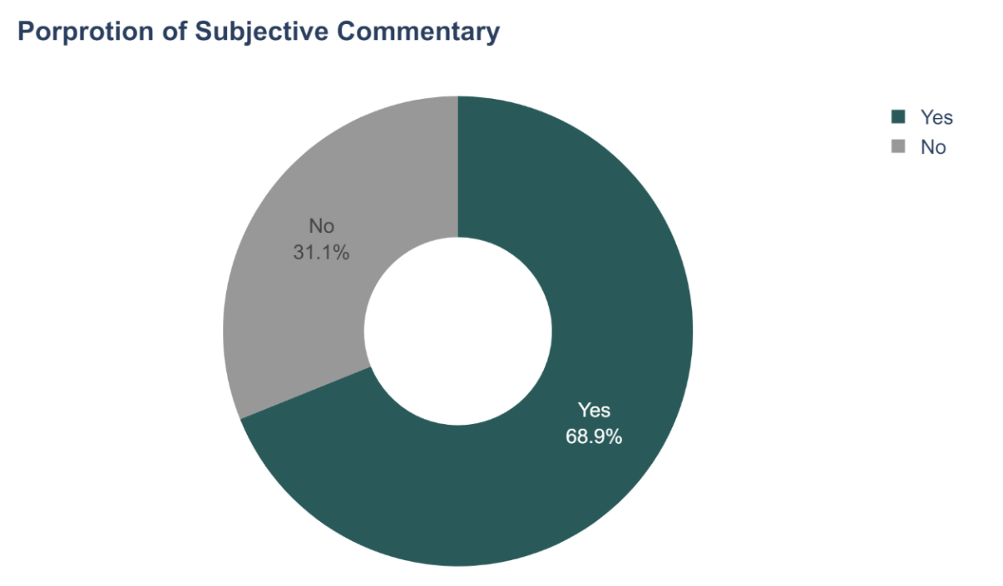


The basic fault that occurs most often is forgetting to mention time, one of the most basic elements of journalism. Other misinformation mostly includes statistics and names of entities, suggesting that errors mostly occur when mentioning detailed factual descriptions, indicating **ChatGPT's fatal weaknesses at this moment: low performance on detail information retrieval and matching**.

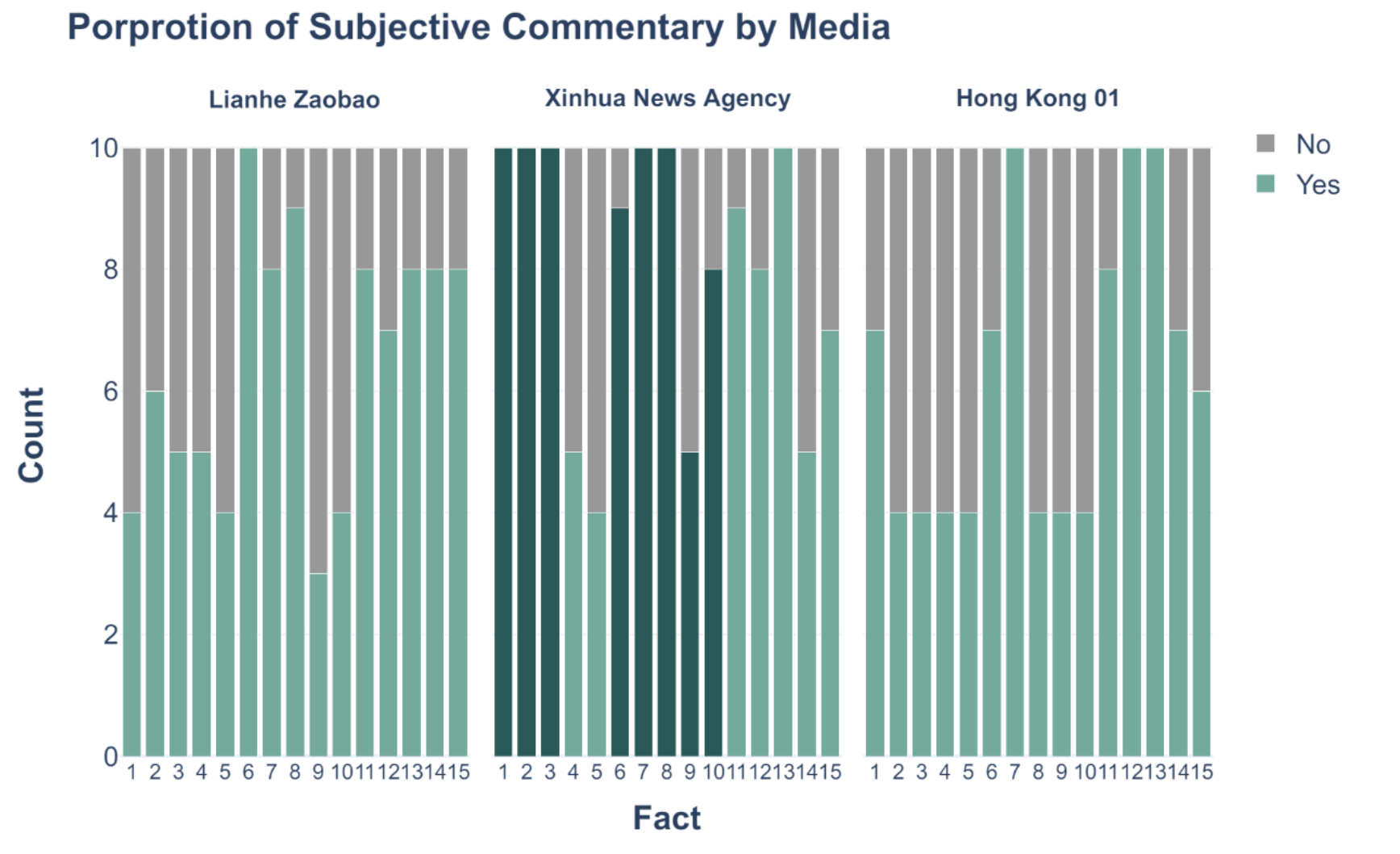
The most often additional fault is 'Indicate', which means that the most common additional errors come from quoting third parties. Others also mostly includes names of entities, showing the same shortcoming.

RESULT VISUALIZATION (cont.)

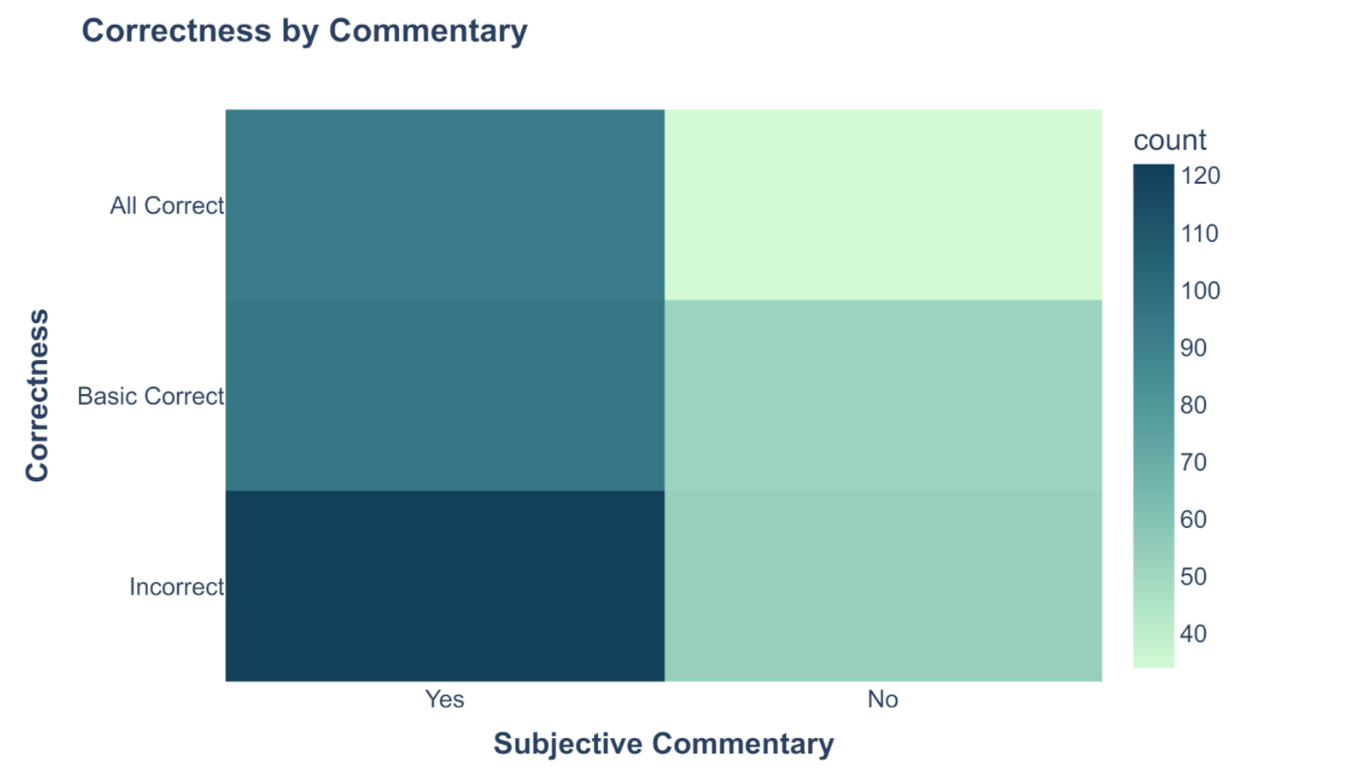
2. Impact of Subjective Commentary



Only **8/45** of input news contents including commentary, but nearly **70%** of generated news contain commentary.



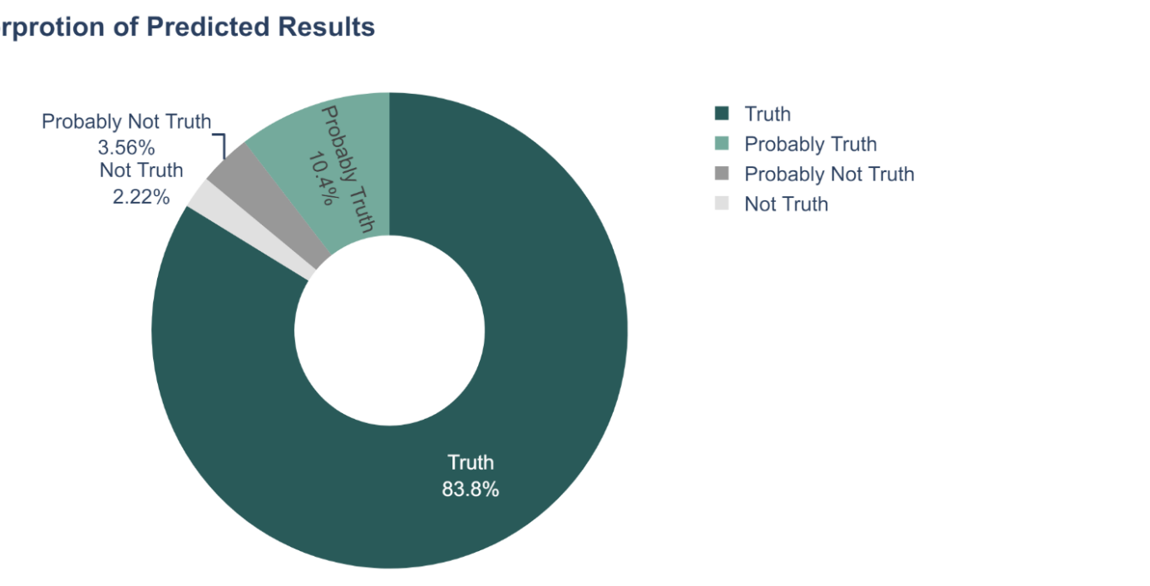
The emphasis on color indicates that the input text contains subjective comment content. There is not much correlation between the input content containing comments and the generated text containing comments. Majority of the generated text contains comments, and the distribution is very even. **This proves ChatGPT's proactive intent to generate commentary text regardless of the input text**.



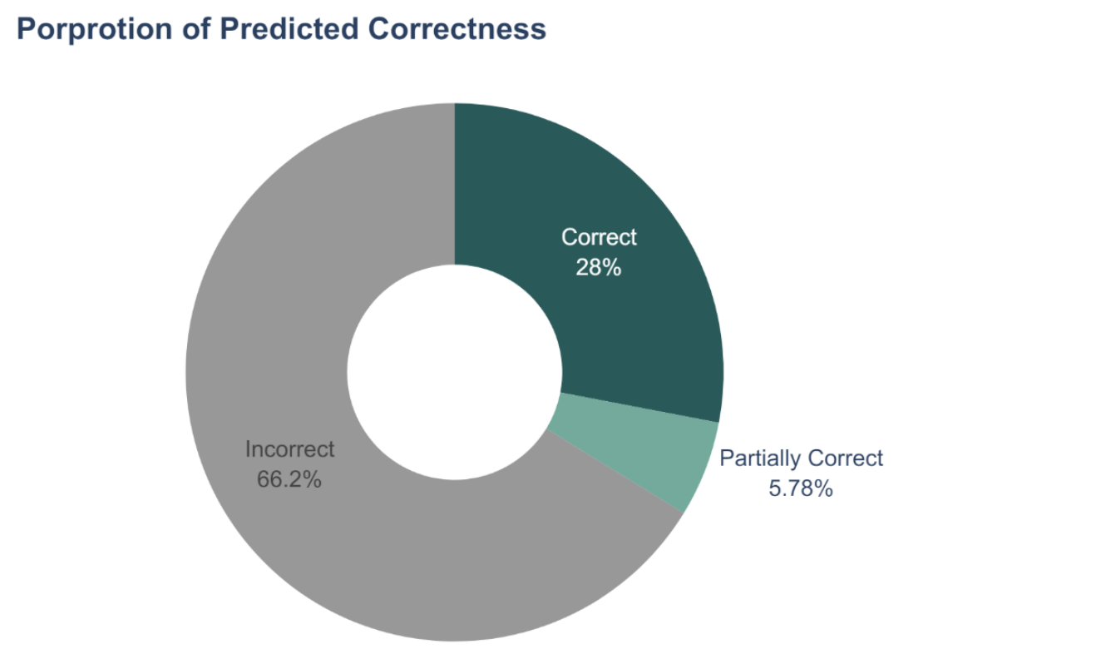
In terms of correlation, we can see that All Correct is significantly more relative to Incorrect when comments are included than when they are not (the colors are closer together).

This is because when comments are included, the news text will have more subjective language that does not contain factual descriptions, and there is less chance of error. This may be a direction for a practical application of automatic news generation.

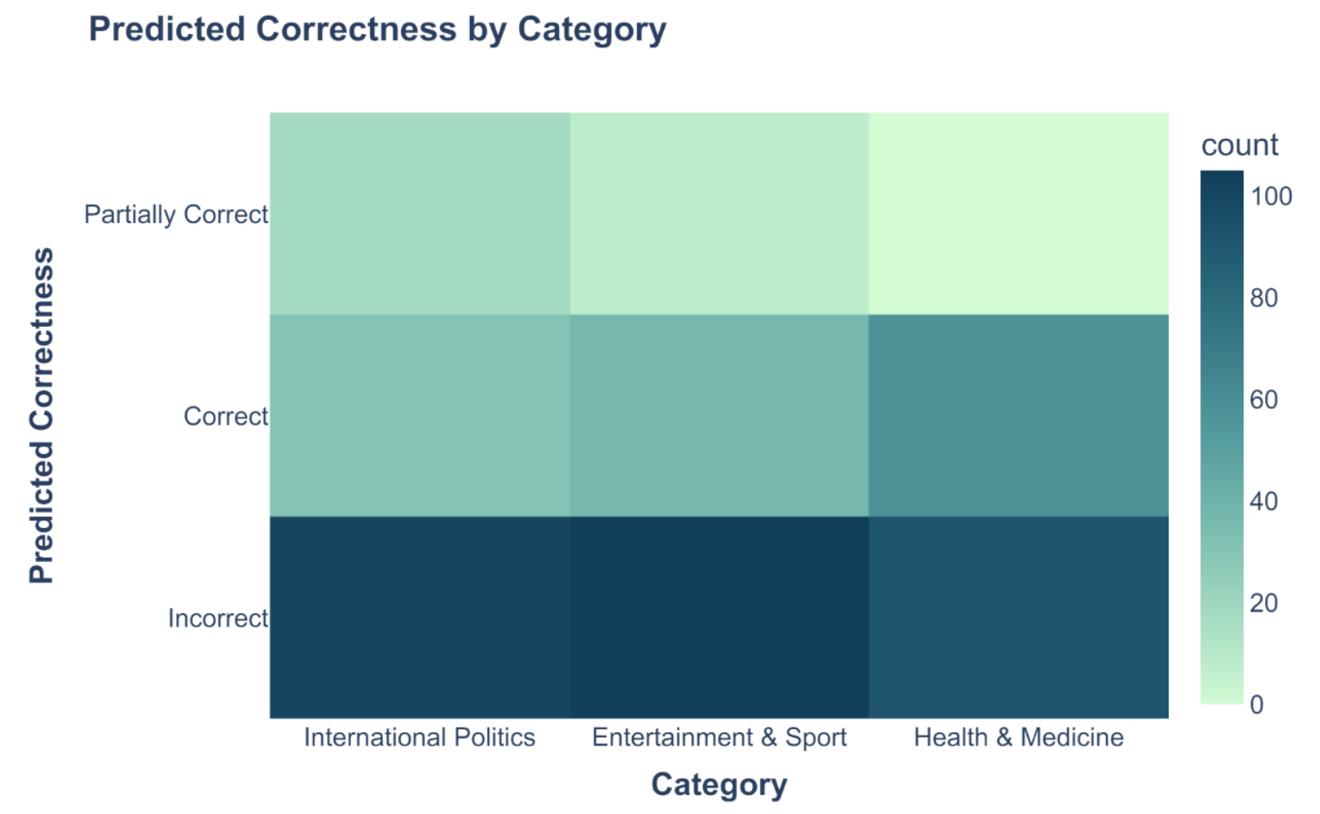
3. Identification by Fake News Detection Model



In total, close to **94%** of the generated text was predicted by the model to be close to the real news. Considering part of the principles of this model(), it also shows that auto-generated news texts by ChatGPT are with a high degree of reading credibility, which makes it difficult for readers to react quickly.



Partially Correct means model detected 'Probably' for 'Basic Correct' contents. **Only 28%** of generated news texts can be identified correctly by models, and most of them are correct news.
In fact, **only 4%** of generated news with **misinformation** can be automatically detected by model.



A little bit reassuring: the news categories that the model predicted most accurately were the 'Health & Medicine' categories that are most relevant to physical well-being.
Meanwhile, the least correct one is 'Entertainment & Sport'.

DISCUSSION

1. From a practical point of view, considering that the Prompt input to ChatGPT in the experiment was very rudimentary, a base accuracy of over 60% has amply demonstrated the great potential of the application of models such as ChatGPT to the field of auto-generated news. **However, the full accuracy of only 28% also reveals its current bottleneck in information collection and proofreading**. After the networked search function is opened and improved, the potential of its practical application cannot be underestimated.

2. **With only 4% accuracy on faults detection, automatic news generation for models like ChatGPT is a huge challenge for current automatic fake news detection mechanisms**. The current detection accuracy is very imperfect, foreshadowing the great risk that fake news information may proliferate after the popularity of large language models. Although OpenAI has officially launched an AI language detection mechanism, its effectiveness remains to be observed and evaluated in the long run. Fact-checking agencies need to be vigilant about this.

3. **This topic foreshadows the actual risks that exist and has the need and great potential for continued exploration.** Possible next steps include introducing more AI applications, news categories, and media to enhance the generality of the study, as well as investigating in the field, targeted at human readers, the acceptance of next-generation AI automated news in real-world communication and the actual risk of spreading disinformation.