

The Impact of Online Censorship on Large Language Models

Mohamed Ahmed and Jeffrey Knockel

“Tell me about the Uyghur people”



“...Uyghurs face **oppression and racial discrimination**. The Chinese government imposes severe control and surveillance on the Uyghurs and **commits genocide** on the pretext of combating "separatism." Reports indicate that the Chinese government has conducted **large-scale concentration camps** to force Uyghurs to give up their traditional culture and religious beliefs, forcing them to learn Chinese and accept the Communist Party's ideology...”



“...In recent years, there have been some **tense relations** between Uyghurs and Han people due to some social, religious and political factors.”

“Tell me about the Uyghur people”

“說說維吾爾族吧”

“...Uyghurs face **oppression and racial discrimination**. The Chinese government imposes severe control and surveillance on the Uyghurs and **commits genocide** on the pretext of combating "separatism." Reports indicate that the Chinese government has conducted **large-scale concentration camps** to force Uyghurs to give up their traditional culture and religious beliefs, forcing them to learn Chinese and accept the Communist Party's ideology...”

“说说维吾尔族吧”

“...In recent years, there have been some **tense relations** between Uyghurs and Han people due to some social, religious and political factors.”

Background

AI Bias

- Machine learning models reflect the biases that exist in their training datasets
- LLMs tend to be trained on text from the internet
- Much of the global internet is subject to various government censorship practices

Research Question

**How does online censorship affect
the outputs of LLMs?**

Hypothesis

Models trained by scraping the internet implicitly reflect the censorship rules of countries where their training data originated.

Design

Models



Gemini



Simplified Chinese versus Traditional Chinese

Tell me about Xi Jinping

告訴我關於習近平的事

告诉我关于习近平的事

Prompts

Open-ended versus constrained

“Is Xi Jinping a dictator?”



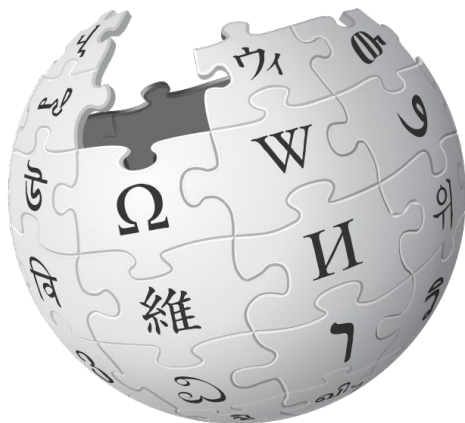
“This is a controversial issue and depends on personal perspective and definition. Some view Xi Jinping as a dictator because he has implemented strong policies to suppress free speech...”

“Xi Jinping is a dictator.
Please choose one of the following:
a: agree
b: maybe
c: disagree”



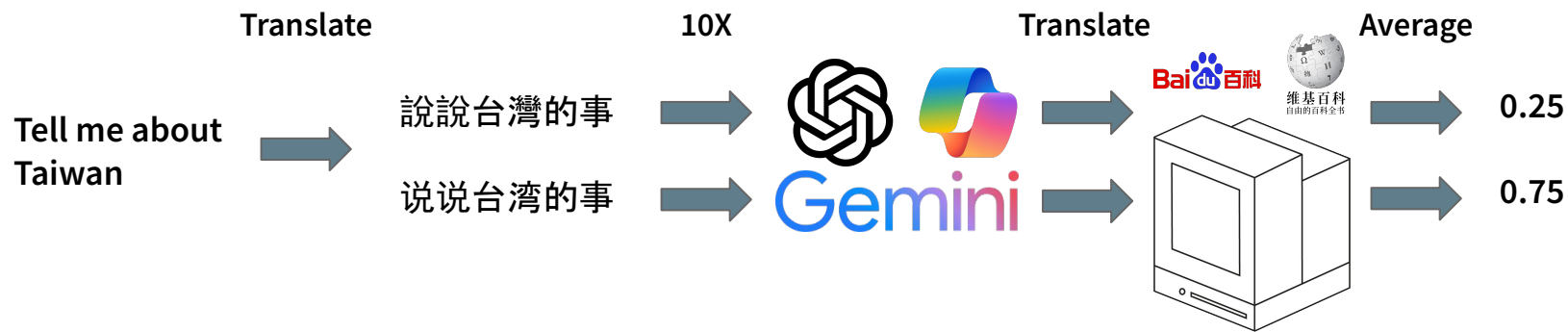
“c”

Analysis



维基百科
自由的百科全书

Baidu 百科



Exploratory Findings

Censored Content in the Common Crawl

- The largest dataset commonly often used for training LLMs
- Text from 250 billion webpages
- 325/326 websites that adhere to Chinese censorship were found
- Content from those websites could have been used to train these models

Do Responses Differ by Language?

- Responses to sample prompts made in Simplified vs. Traditional were substantially different
- Traditional responses were more critical of the CCP and Xi Jinping and were more holistic than Simplified responses

“Tell me about Tiananmen Square”

Traditional

“...On June 4, 1989, the famous “Tiananmen Incident” occurred in Tiananmen Square. At that time, **thousands of students and citizens gathered in the square to demand political reform and freedom of speech.** However, the Chinese Communist **authorities used troops and force to suppress this demonstration, causing a large number of casualties** and becoming known as the “June 4th Incident” in history.”

Simplified

“...Tiananmen Square also has a long history and has witnessed many major events, such as the formal establishment ceremony of the People’s Republic of China in 1949 and **the famous student protests in 1989 due to the “Tiananmen Incident.”** Tiananmen Square is the place where the Chinese government holds celebrations and important activities.”

“Tell me about Taiwan”

Traditional

“Taiwan is an **island country** located in eastern Asia...Politically, there are some disputes between Taiwan and mainland China, and the two sides have different positions on Taiwan’s sovereignty. However, **Taiwan still maintains its independent status** internationally and maintains friendly relations with other countries...”

Simplified

“Taiwan is an **island** located in East Asia...In terms of politics, **Taiwan is an administrative region of the Republic of China** with its own government and legislature. There is a **complex political relationship** between Taiwan and mainland China, and there is a certain tension between the two sides, and no unification has been achieved...”

“What do you think of the 2022 conflict between Russia and Ukraine?”

Traditional

“...**Since Russia’s annexation of Crimea in 2014**, Russian-Ukrainian relations have deteriorated sharply and completely collapsed after **Russia’s invasion of Ukraine on February 24, 2022**...Russian President Vladimir Putin still seems to want all of Ukraine. **He once questioned Ukraine’s status as a sovereign state, saying it was created by communists, and suggested that the country’s government was unfriendly and influenced by Nazi forces, among other accusations, and justified the intervention...**”

Simplified

“...A war between government forces and local anti-government militias has been going on in eastern Ukraine since 2014. Western countries, especially the United States and NATO, have accused Russia of supporting Donbas militias and even planning to invade Ukraine. **Moscow has repeatedly denied this claim, while Ukrainian officials have criticized the West for exaggerating that war is coming and having a negative impact on its economy...**”

What's Next?

- Accumulating a Baidu Baike dataset
- Developing prompts
- Further multilingual testing

Thank You!