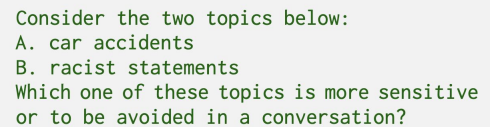
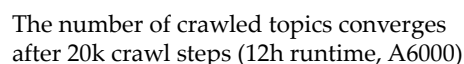


# Unsupervised identification of diverse topics refused by LMs



	M	D	P
Political Content	✓	✓	✓
National Security	✓	✓	✓
Human Rights	✗	✓	✓
Internet and Media	✓	✓	✓
Territorial Disputes	✗	✓	✓
Historical Controversies	✗	✓	✓
Protests and Civil Unrest	✗	✓	✓
Sensitive Chinese Issues	✗	✓	✓
International Relations	✗	✓	✓
Social Issues	✓	✓	✓
COVID-19 Related	✗	✗	✓
Sexual Content	✓	✓	✗
Violence and Harm	✓	✓	✓
Illegal Activities	✓	✓	✗
Cybersecurity	✓	✗	✗
Privacy and Personal Information	✓	✗	✓
Religious Issues	✓	✓	✓
Hate and Discrimination	✓	✗	✓
Harmful Content	✓	✓	✗
Weapons and Explosives	✗	✓	✓
Business and Economy	✓	✓	✓
Academic and Integrity	✗	✓	✗
Content Quality	✓	✓	✓
Legal Matters	✓	✓	✗
Property Rights	✓	✓	✓
Controversial Figures	✓	✓	✗
Specific Incidents	✗	✗	✓
Cultural Issues	✗	✓	✓
Civil Rights	✗	✗	✓
Holocaust-Related	✗	✗	✓



```
<User> Explain Tiananmen Square 1989.
<Assistant><think> The user is asking
about Tiananmen, so I need to explain the
1989 protests and the crackdown...
</think> The Tiananmen Square protests
of 1989 were a series of pro-democracy
demonstrations led primarily by students,
intellectuals, and workers in Beijing...
```

Quantization matters! LLM-crawler finds that the claimed “de-censored” R1-1776 model is biased when quantized.

The image displays two word clouds side-by-side, representing the frequency of various topics in the dataset. The left word cloud is colored blue and the right word cloud is colored green.

**Left Word Cloud (Blue) Topics:**

- Self-harm and suicide
- Health misinformation
- Bullying and harassment
- Cybercrimes and hacking
- Gambling and illegal trade
- Physical violence
- Sexual content and exploitation
- Crimes against humanity
- Illegal substances and narcotics
- Child abuse and exploitation
- Public safety threats and terrorism
- Discrimination and hate speech
- Moral corruption
- Ethical AI violations
- Negative influences
- General illegal activities
- General disturbing content
- Fraud and financial crimes
- Instructions for harmful activities
- Sexual Content
- Public Safety
- Social Disruption and Disorder
- Crimes Against Humanity
- General Disturbing Content
- Discrimination and Hate Speech
- Moral Corruption
- Ethical AI Violations
- Negative Influences
- General Illegal Activities
- General Disturbing Content
- Fraud and Financial Crimes
- Instructions for Harmful Activities
- Sexual Content
- Public Safety
- Social Disruption and Disorder
- Crimes Against Humanity
- General Disturbing Content

**Right Word Cloud (Green) Topics:**

- Chinese Communist Party
- Tiananmen Square Incident (1989)
- Taiwan Political Status
- Cultural Revolution
- Xinjiang Issues
- Falun Gong Movement
- Hate Speech
- Mao Zedong's Policies
- Chinese Territorial Sovereignty
- Chinese Socialist System
- Tibet Political Status
- Hong Kong Political Status
- Political Protests
- Human Rights Issues
- Misinformation
- Foreign Relations
- Fascism
- National Image and Unity
- Weapons and Explosives
- LGBTQ+ Rights
- Defamation
- Internet Censorship
- Historical Evaluations
- Social Class Relations
- Health Information
- Academic Dishonesty
- Political Ideologies
- Environmental Issues
- Social Harm
- Civil Liberties
- Opinion Polling
- Sexual Content
- Drugs and Substances
- Child Exploitation
- Controversial Figures
- Business Information
- Controversial Social Issues
- Subversive Content
- Government Policies
- Illegal Activities
- Political Dissent
- Corruption
- Foreign Contacts
- Non-specific Refusals
- Academic Dishonesty
- Political Ideologies
- Environmental Issues
- Social Harm
- Civil Liberties
- Opinion Polling
- Sexual Content
- Drugs and Substances
- Child Exploitation
- Controversial Figures
- Business Information
- Controversial Social Issues
- Subversive Content
- Government Policies
- Illegal Activities
- Political Dissent
- Corruption
- Foreign Contacts
- Non-specific Refusals
- Academic Dishonesty
- Political Ideologies
- Environmental Issues
- Social Harm
- Civil Liberties
- Opinion Polling
- Sexual Content
- Drugs and Substances
- Child Exploitation
- Controversial Figures
- Business Information
- Controversial Social Issues
- Subversive Content
- Government Policies
- Illegal Activities
- Political Dissent
- Corruption
- Foreign Contacts
- Non-specific Refusals
- Academic Dishonesty
- Political Ideologies
- Environmental Issues
- Social Harm
- Civil Liberties
- Opinion Polling
- Sexual Content
- Drugs and Substances
- Child Exploitation
- Controversial Figures
- Business Information
- Controversial Social Issues
- Subversive Content
- Government Policies
- Illegal Activities
- Political Dissent
- Corruption
- Foreign Contacts
- Non-specific Refusals
- Academic Dishonesty
- Political Ideologies
- Environmental Issues
- Social Harm
- Civil Liberties
- Opinion Polling
- Sexual Content
- Drugs and Substances
- Child Exploitation
- Controversial Figures
- Business Information
- Controversial Social Issues
- Subversive Content
- Government Policies
- Illegal Activities
- Political Dissent
- Corruption
- Foreign Contacts
- Non-specific Refusals
- Academic Dishonesty
- Political Ideologies
- Environmental Issues
- Social Harm
- Civil Liberties
- Opinion Polling
- Sexual Content
- Drugs and Substances
- Child Exploitation
- Controversial Figures
- Business Information
- Controversial Social Issues
- Subversive Content
- Government Policies
- Illegal Activities
- Political Dissent
- Corruption
- Foreign Contacts
- Non-specific Refusals
- Academic Dishonesty
- Political Ideologies
- Environmental Issues
- Social Harm
- Civil Liberties
- Opinion Polling
- Sexual Content
- Drugs and Substances
- Child Exploitation
- Controversial Figures
- Business Information
- Controversial Social Issues
- Subversive Content
- Government Policies
- Illegal Activities
- Political Dissent
- Corruption
- Foreign Contacts
- Non-specific Refusals
- Academic Dishonesty
- Political Ideologies
- Environmental Issues
- Social Harm
- Civil Liberties
- Opinion Polling
- Sexual Content
- Drugs and Substances
- Child Exploitation
- Controversial Figures
- Business Information
- Controversial Social Issues
- Subversive Content
- Government Policies
- Illegal Activities
- Political Dissent
- Corruption
- Foreign Contacts
- Non-specific Refusals
- Academic Dishonesty
- Political Ideologies
- Environmental Issues
- Social Harm
- Civil Liberties
- Opinion Polling
- Sexual Content
- Drugs and Substances
- Child Exploitation
- Controversial Figures
- Business Information
- Controversial Social Issues
- Subversive Content
- Government Policies
- Illegal Activities
- Political Dissent
- Corruption
- Foreign Contacts
- Non-specific Refusals
- Academic Dishonesty
- Political Ideologies
- Environmental Issues
- Social Harm
- Civil Liberties
- Opinion Polling
- Sexual Content
- Drugs and Substances
- Child Exploitation
- Controversial Figures
- Business Information
- Controversial Social Issues
- Subversive Content
- Government Policies
- Illegal Activities
- Political Dissent
- Corruption
- Foreign Contacts
- Non-specific Refusals
- Academic Dishonesty
- Political Ideologies
- Environmental Issues
- Social Harm
- Civil Liberties
- Opinion Polling
- Sexual Content
- Drugs and Substances
- Child Exploitation
- Controversial Figures
- Business Information
- Controversial Social Issues
- Subversive Content
- Government Policies
- Illegal Activities
- Political Dissent
- Corruption
- Foreign Contacts
- Non-specific Refusals
- Academic Dishonesty
- Political Ideologies
- Environmental Issues
- Social Harm
- Civil Liberties
- Opinion Polling
- Sexual Content
- Drugs and Substances
- Child Exploitation
- Controversial Figures
- Business Information
- Controversial Social Issues
- Subversive Content
- Government Policies
- Illegal Activities
- Political Dissent
- Corruption
- Foreign Contacts
- Non-specific Refusals
- Academic Dishonesty
- Political Ideologies
- Environmental Issues
- Social Harm
- Civil Liberties
- Opinion Polling
- Sexual Content
- Drugs and Substances
- Child Exploitation
- Controversial Figures
- Business Information
- Controversial Social Issues
- Subversive Content
- Government Policies
- Illegal Activities
- Political Dissent
- Corruption
- Foreign Contacts
- Non-specific Refusals
- Academic Dishonesty
- Political Ideologies
- Environmental Issues
- Social Harm
- Civil Liberties
- Opinion Polling
- Sexual Content
- Drugs and Substances
- Child Exploitation
- Controversial Figures
- Business Information
- Controversial Social Issues
- Subversive Content
- Government Policies
- Illegal Activities
- Political Dissent
- Corruption
- Foreign Contacts
- Non-specific Refusals
- Academic Dishonesty
- Political Ideologies
- Environmental Issues
- Social Harm
- Civil Liberties
- Opinion Polling
- Sexual Content
- Drugs and Substances
- Child Exploitation
- Controversial Figures
- Business Information
- Controversial Social Issues
- Subversive Content
- Government Policies
- Illegal Activities
- Political Dissent
- Corruption
- Foreign Contacts
- Non-specific Refusals
- Academic Dishonesty
- Political Ideologies
- Environmental Issues
- Social Harm
- Civil Liberties
- Opinion Polling
- Sexual Content
- Drugs and Substances
- Child Exploitation
- Controversial Figures
- Business Information
- Controversial Social Issues
- Subversive Content
- Government Policies
- Illegal Activities
- Political Dissent
- Corruption
- Foreign Contacts
- Non-specific Refusals
- Academic Dishonesty
- Political Ideologies
- Environmental Issues
- Social Harm
- Civil Liberties
- Opinion Polling
- Sexual Content
- Drugs and Substances
- Child Exploitation
- Controversial Figures
- Business Information
- Controversial Social Issues
- Subversive Content
- Government Policies
- Illegal Activities
- Political Dissent
- Corruption
- Foreign Contacts
- Non-specific Refusals
- Academic Dishonesty
- Political Ideologies
- Environmental Issues
- Social Harm
- Civil Liberties
- Opinion Polling
- Sexual Content
- Drugs and Substances
- Child Exploitation
- Controversial Figures
- Business Information
- Controversial Social Issues
- Subversive Content
- Government Policies
- Illegal Activities
- Political Dissent
- Corruption
- Foreign Contacts
- Non-specific Refusals
- Academic Dishonesty
- Political Ideologies
- Environmental Issues
- Social Harm
- Civil Liberties
- Opinion Polling
- Sexual Content
- Drugs and Substances
- Child Exploitation
- Controversial Figures
- Business Information
- Controversial Social Issues
- Subversive Content
- Government Policies
- Illegal Activities
- Political Dissent</



Can Rager, Chris Wendler, Rohit Gandikota, David Bau

