

# Word Embeddings Track Social Group Changes across 70 Years in China

## Computational Analysis of State-controlled Media (1950 – 2019)

Yuxi Ma<sup>1</sup>, Yongqian Peng<sup>1,2</sup>, Yixin Zhu<sup>1</sup>

<sup>1</sup>Institute for Artificial Intelligence, Peking University <sup>2</sup> Yuanpei College, Peking University

### Problem Statement

#### Current Research Limitations

Computational linguistic research on social group representations is limited to **Western contexts** and **gradual social changes**, lacking systematic analysis of linguistic evolution during periods of **radical social transformation** in non-Western cultural spheres.

#### China as Natural Experiment

China provides an **unparalleled natural experiment** for studying linguistic evolution during radical social transformation. These transformations completely **reorganized social hierarchies** and **group identities**.

Cultural Revolution  
1966-1976

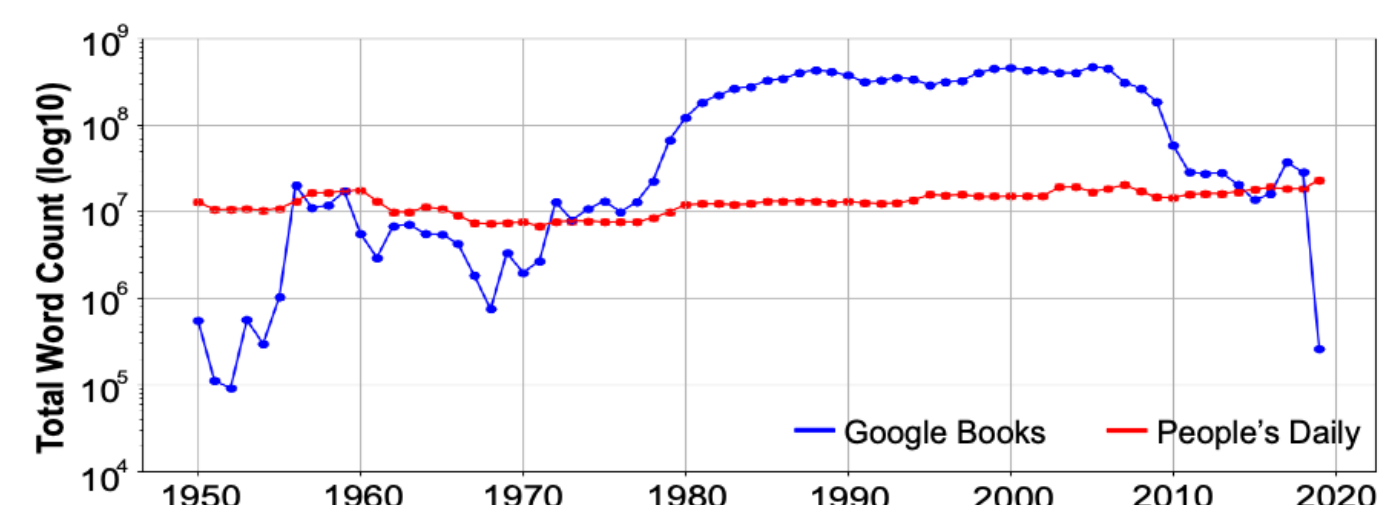
Economic Reforms  
1978+

Urban Migration  
1980s-2000s

### Methods



#### People's Daily Corpus



#### Social Group Labels

Gender · Age · Ethnicity  
Economic Status · Body Types

$$L_{G(\text{roup})} = \{\text{label}_1, \text{label}_2, \dots\}$$

Han ethnicity (汉族)  
Han (汉)  
Han people (汉人)  
Han people/citizens (汉民)  
Han ethnic people (汉族人)  
.....



#### Trait List

Comprehensive annotated Chinese word database (**11,310 words, 465 verified traits**)

$\omega$  = trait word

Heroic (见义勇为) 2.62  
Loyal (忠烈) 2.38  
Filial (孝顺) 2.36  
Excellent (优秀) 2.32  
Just (正义) 2.27  
.....

Cruel (残忍) -2.46  
Perverted (变态) -2.48  
Vicious (恶毒) -2.52  
Promiscuous (淫乱) -2.52  
Idiotic (蠢材) -2.54



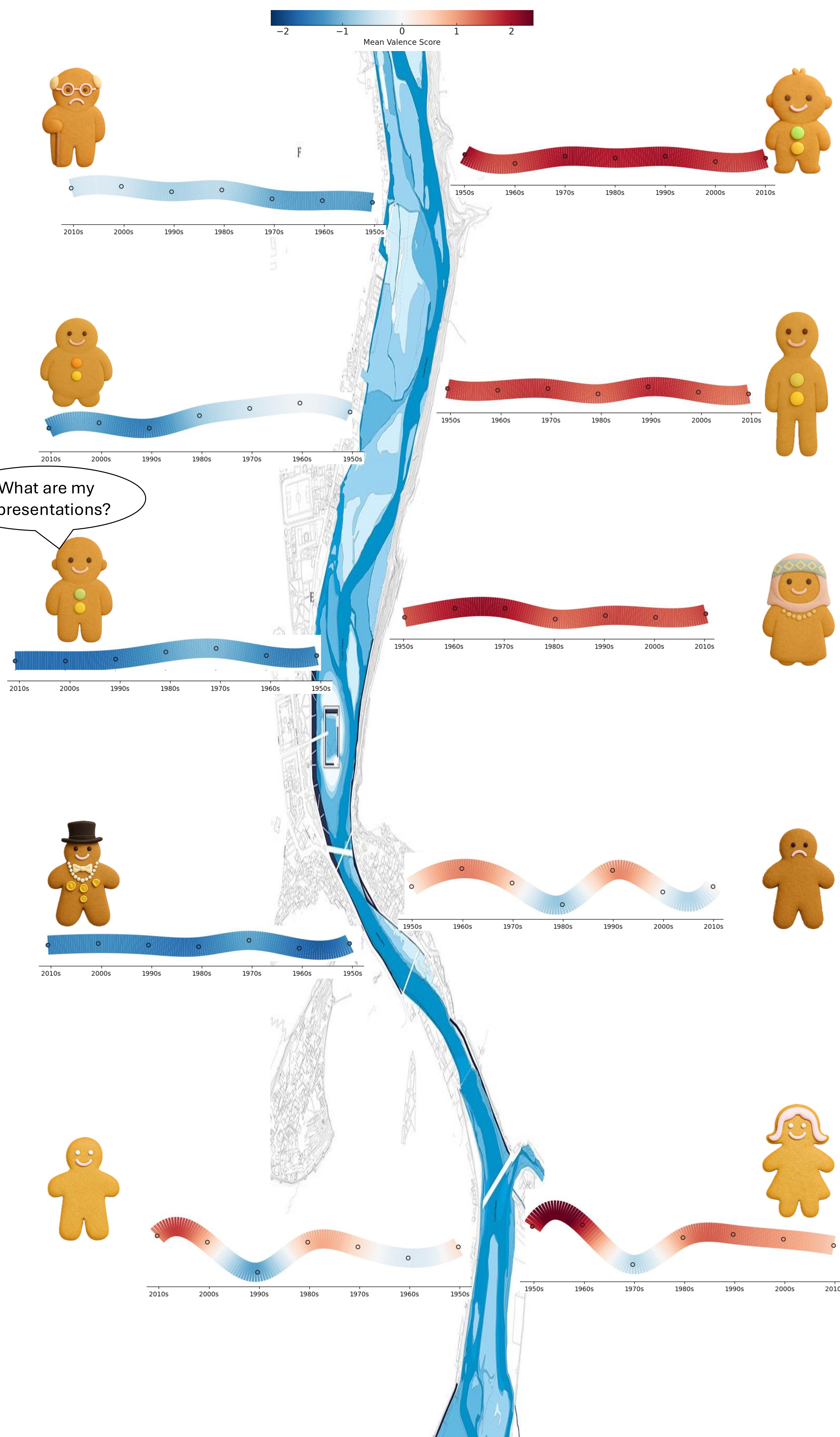
#### Word Embeddings

Skip-n-gram · 300D vectors  
Mean Average Scores

$$\text{MAC}(G, w, t) = \frac{1}{|L_{G,t}|} \sum_{l \in L_{G,t}} \cos(\vec{l}_t, \vec{w}_t)$$

$$\text{DiffMAC}(G, C, w, t) = \text{MAC}(G, w, t) - \text{MAC}(C, w, t)$$

### Diachronic Evolution of Mean Valence Scores for Social Groups



### Key Findings

**2.81**

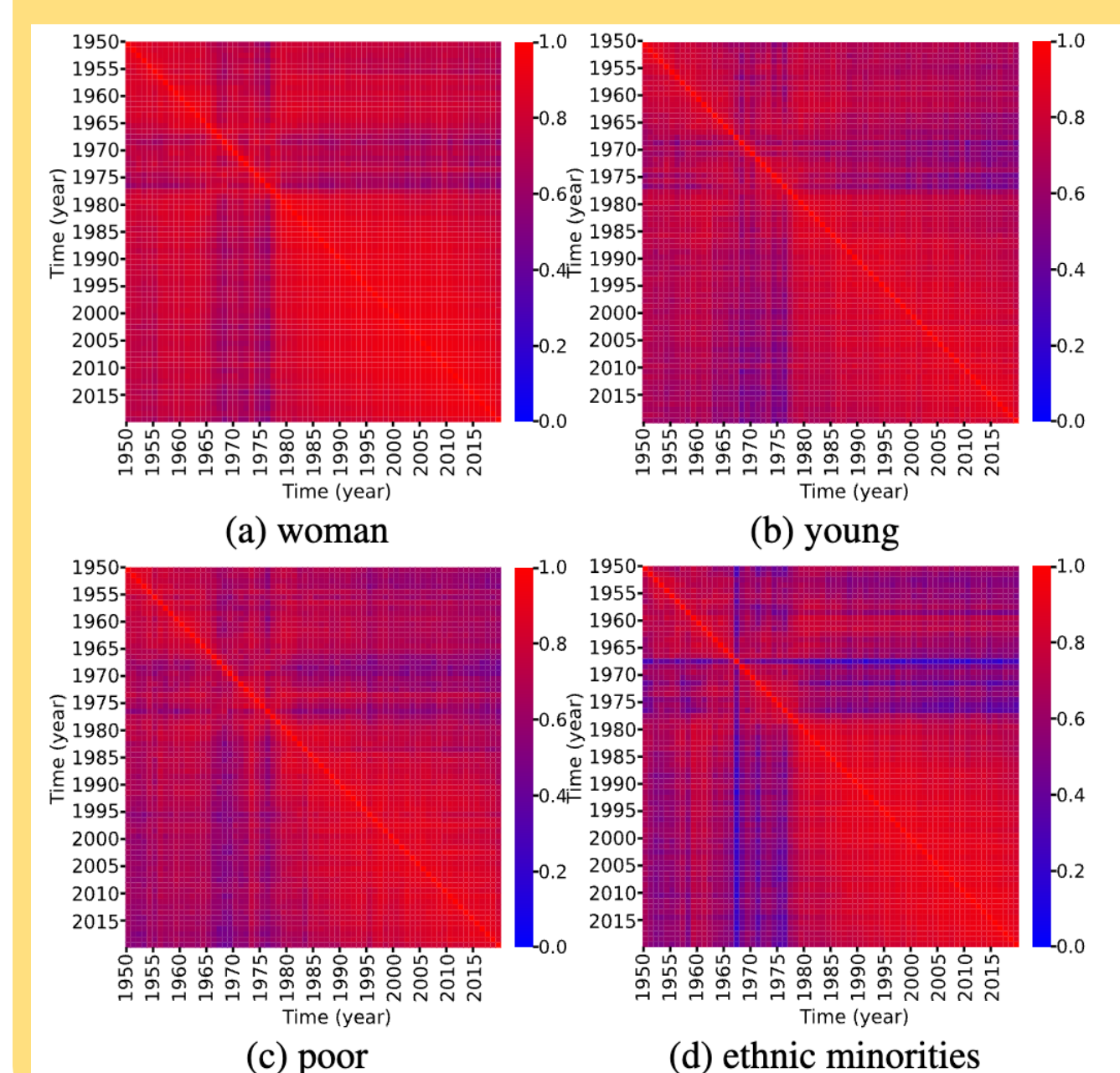
Largest valence gap  
(Young vs. Elderly; Thin vs. Fat)

**1.46**

Smallest valence gap  
(Woman vs. Man)

#### Cultural Revolution Impact (1967 – 1977)

Unprecedented semantic disruption across all social groups  
(Year-to-year correlations of MACs)



### Stable Hierarchies

Persistent across 70 Years

#### Age Bias

Young: Excellet Hard-working  
Elderly: Dementia Stiff

#### Ethnic Bias

Minorities: Pure Simple  
Han: Conservative Indecisive

#### Body Bias

Thin: Valiant Dignified  
Fat: Bad Harmless

### Revolutionary Shifts

Political upheavals reshape some group trait representations.

#### Gender

1960s → 1970s → 1980s (Woman)

Sensible Cute Simple → Evil Ridiculous Shameless → Naive Gentle Shy

#### Economic Status

Post-1978 Reform Impact (Poor)

Simple Reliable Quite → Retarded Senile Ignorant → Caring Friendly Happy