Racial Representation in US Cinema, 1910–2020

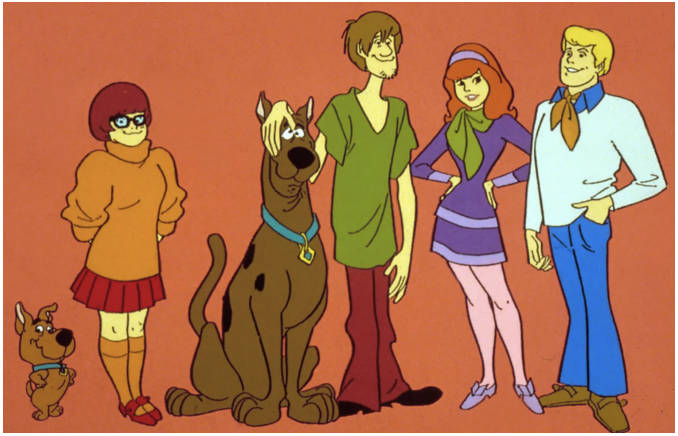
Raphaël Boulat (Bocconi) & Alvise Scarabosio (Stanford)

La Strada Seminar

April 24, 2026

Motivation - Early Exposure to Representation

Figure: Daphne is Ginger!



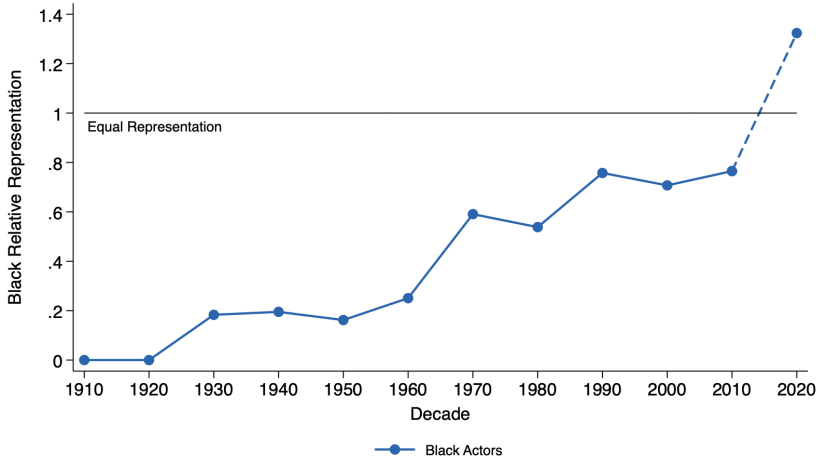
Motivation - Beyond Anecdotal Evidence

- Role models in media impact beliefs, aspirations, and behavior (Riley '24; LaFerrara '12, '16: Breda et al. '18)

- Culture as a residual: physical capital, human capital and institutions leave a large residual when explaining differences in economic outcomes across groups, individuals, and countries
  ↪ open the "black box of culture"

- The global movie and video market reached $328 billion in 2025 is projected to grow to $418 billion by 2029 (Source)

Motivation - A Century of Change in On-Screen Representation



This Paper

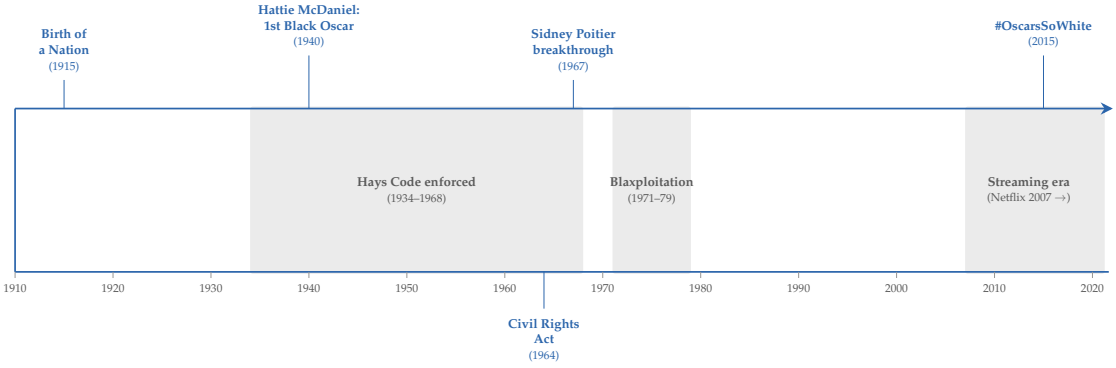
- Provides descriptive evidence on Black on-screen representation over a century
- Goes beyond mere presence and analyzes movie scripts to understand the importance and social positioning of Black characters
- Attempts to explain the changing patterns in racial representation by investigating the link between movie popularity and racial attitudes

Literature

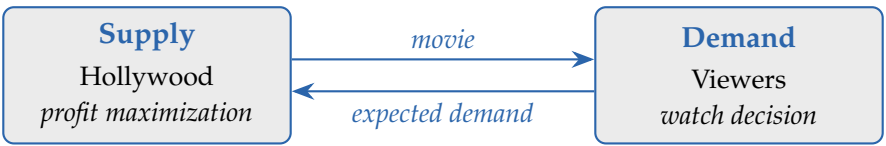
1. **Measuring Racial Representation in Movies**
   - Computer science work on character identification and demographic inference (e.g., Ramakrishna et al. '17, Guha et al. '15, Somandepalli et al. '21)
   - Evidence scattered across studies, genres, and short time windows
   - **This paper:** a unified, century-long picture of racial representation in Hollywood (1910–2020), with a deeper look at how groups are portrayed (e.g., occupational and criminal roles)

2. **Movies, Race, and Attitudes**
   - Movies as cultural products shaped by and shaping audiences (Michalopoulos & Rauh '25)
   - Movies as drivers of racial attitudes and violence (Ang '23; Esposito et al. '23)
   - Demand for representation in cultural consumption (Adukia et al. '23)
   - **This paper:** links local political attitudes to interest in movies with varying racial representation

Historical Background



Conceptual Framework



A two-sided market: studios produce movies anticipating demand; viewers choose what to watch based on what is supplied.

Data creation - Images and Race

- Dataset of $\sim$3000 movies (1910-2023)
- Created match **actor - character - script**
- Scraped actor images from IMDB
- Predicted actor's race using computer vision (Molmo-7B) and actor's skin-tone ($n \approx 37000$)
- Compare performance with 1800 hand-labeled actor pictures

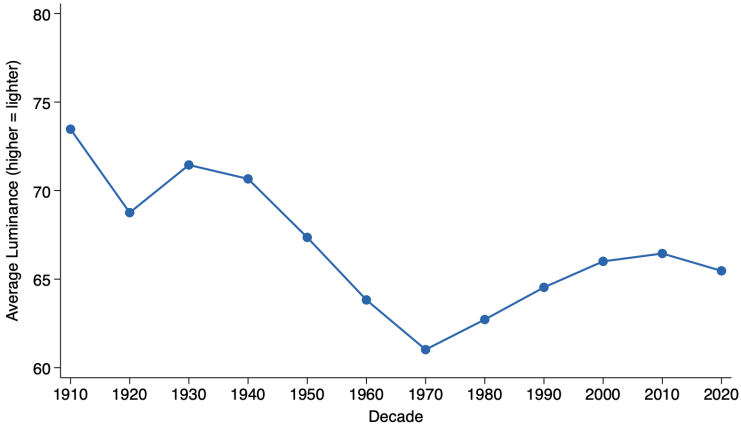
▸ Accuracy ▸ Precision ▸ Recall ▸ F1 ▸ Image Pipeline

Data creation - Scripts

- Pre-processing: OCR quality filter (min. 60% real English words) + decade-stratified sampling to ensure temporal balance → pilot sample of 450 movies from 1950 to 2019.
- Structuring: Gemini 2.5 Flash-Lite annotates every line with one of 12 structural labels (*dialogue*, *action*, *scene_heading*, ...) using chunked prompting with boundary context
- Actor matching: speaker cues linked to actors via fuzzy Levenshtein matching — yields per-character dialogue tied to race & skin-tone predictions
- Output: character-level dialogue dataset merged with image-based race and skin-tone predictions

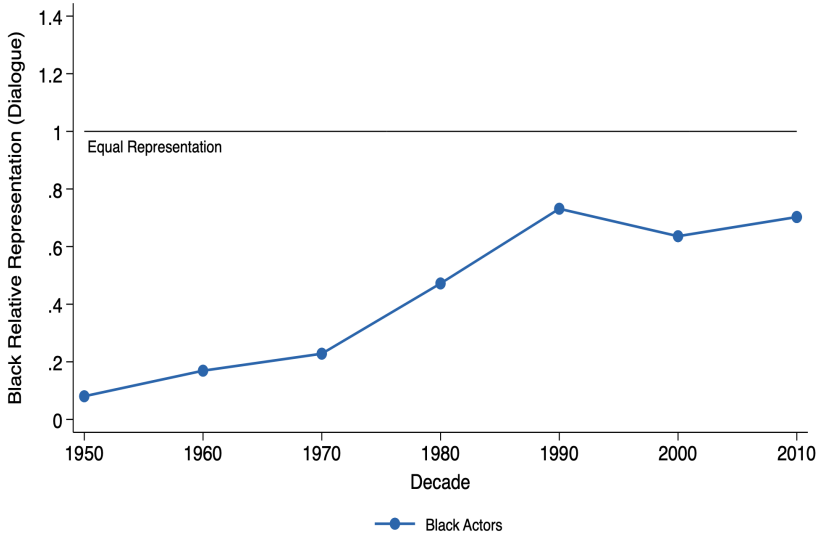
▸ Text Pipeline ▸ Gemini Prompt

Skin Tone Over Time



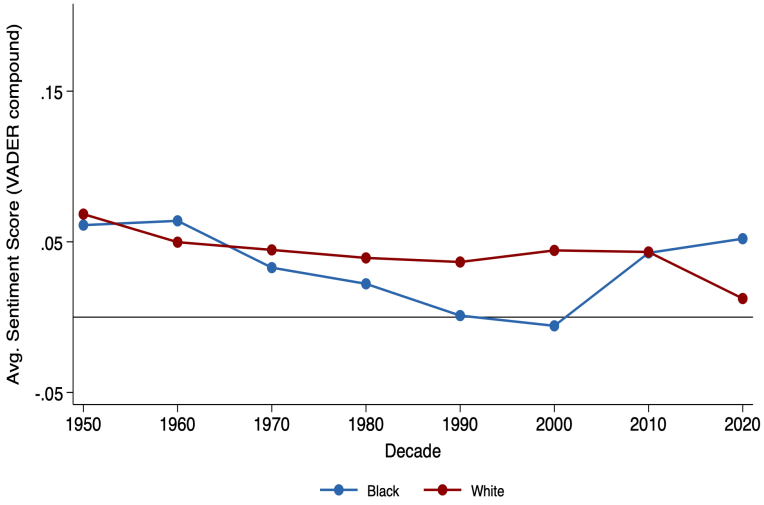
▸ Distributions ▸ Blacks only

Dialogue Over Time

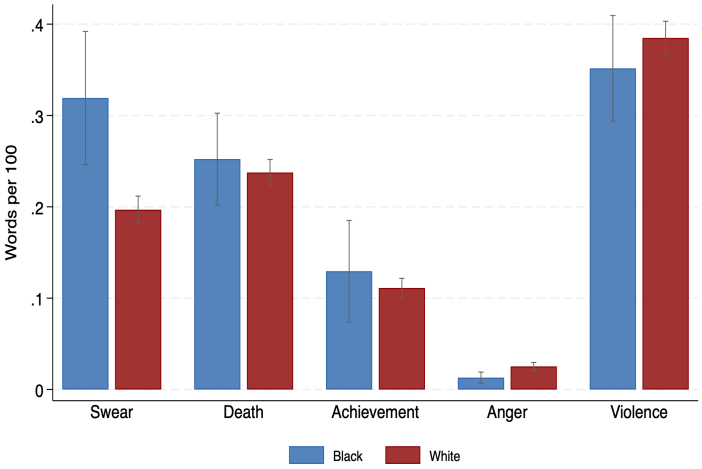


▸ Lead Roles

Sentiment Analysis Averages Over Time By Race



Sentiment Analysis: LIWC



▸ Details

Next steps: Criminal Representation

We could try to classify narrative role relative to crime:

- Perpetrator (commits crimes, leads criminal enterprise)
- Enforcer (police, detective,...)
- Victim (targeted by crime)
- Bystander (no particular relation to crime in the script)

Note: for now, we are using data with dialogue only. One thing that could be interesting in the future is to measure interactions. Which character is committing crime on whom.

- Is it within race violence?
- Are Blacks depicted as a danger for whites?

▸ Gemini Prompt

Next steps: More Text Analysis + Comparison to Posters

- Character specific NLP (occupation, ...)
- Interactions and narratives (probably need more than simply the dialogue)
- **Compare visual representation on posters vs. script representation (e.g. share of words vs share of space $\rightarrow$ marketing angle**

Political attitude and movie success: a demand story?

- Earlier slides documented changing patterns in racial representation. *Why?*.
- Michaelopoulos & Rauh '25 explain that movies closer to a society's tradition are more likely to be screened and generate high revenue ⇒ evidence that *consumer demand*, not just studio supply, shapes movie success
- What about demand for racial representation in the US?
  - **Question:** Do more conservative regions show differential search interest in films based on their racial representation?
  - **Challenge:** Movie-level revenue data is private/for profit at the DMA level.
  - **Solution:** Match Google Trends Data with CCES data on racial attitudes.

Specification

$$\log(I_{ij}) = \delta_i + \delta_j + \theta(\mathcal{R}_i \cdot P_j) + \varepsilon_{ij}$$

where

- I_{ij} is Google Trends search intensity index for movie i in DMA j (averaged over a 1-year afer release)
- $\mathcal{R}_i$ is the movie's racial representation measure (share of black actors in the movie)
- P_j is DMA-level racial progressiveness constructed as the weighted mean across respondents of a 6-item CCES racial-attitudes scale
- θ is the coefficient of interest: captures how steeply the interaction covaries with $\log(I_{ij})$ across DMAs, as a function of $\mathcal{R}_i$.

▸ Racial Progressiveness

Google Trends Data

What Trends measures. For movie i in DMA j over a fixed time window, define the search share

$$s_{ij} \equiv \frac{\#\text{searches for } i \text{ in } j}{\#\text{total searches in } j}.$$

This automatically controls for DMA size: small markets are not penalized for having fewer total searches.

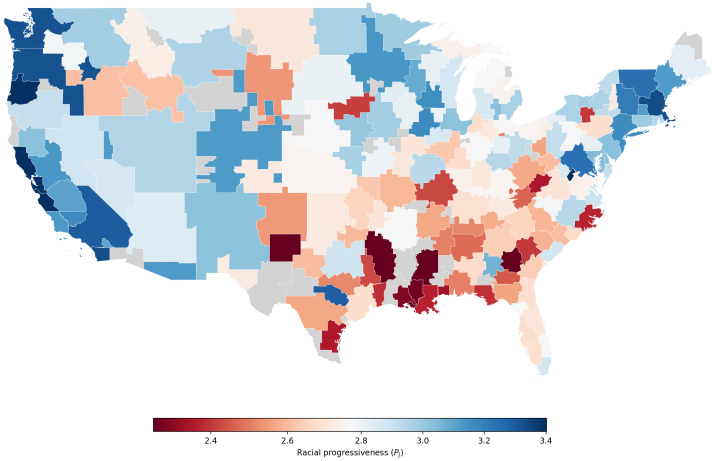
What Trends reports. Not s_{ij} itself, but a per-movie normalized index. Let $\tilde{s}_i = \max_j\{s_{ij}\}$ be the share in movie i's peak DMA. Trends returns

$$I_{ij} \equiv 100 \cdot \frac{s_{ij}}{\tilde{s}_i} \in [0, 100].$$

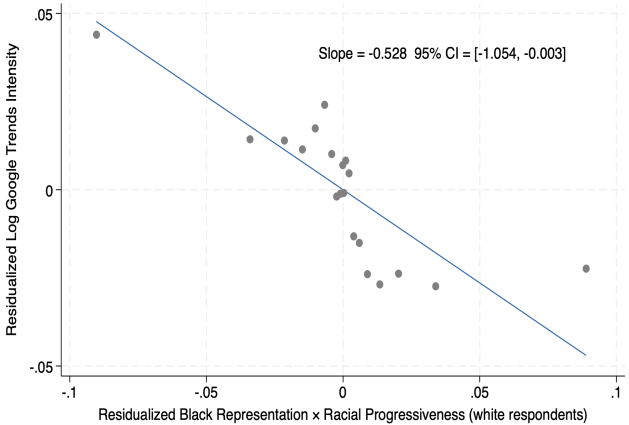
The peak DMA gets 100; every other DMA is scaled relative to it.

▸ Log Specification Justification ▸ Identify Google Trends

Racial Progressiveness (P_j) - Non-Hispanic Whites

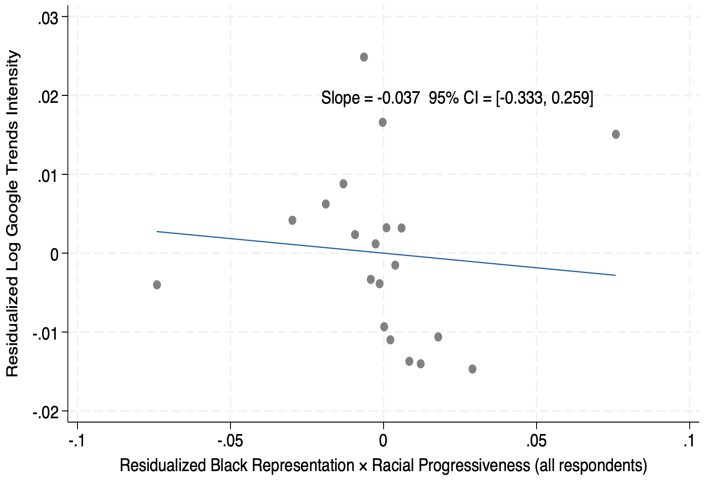


Preliminary Evidence - Non-Hispanic Whites

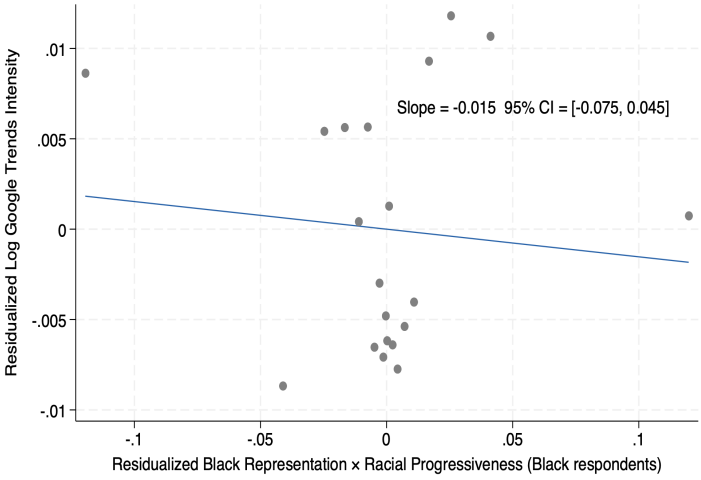


▸ Average DMA gradient

Preliminary Evidence II - All Respondents



Preliminary Evidence III - Black Respondents



Interpreting the Negative Slope

- **The puzzle:** Within-movie, within-DMA, θ is *negative*: more racially progressive DMAs over-search *less* for high representatives movies than for low representatives movies.

- **Potential Interpretation:** Attention $\neq$ demand. Google Trends captures search interest, which mixes enthusiasm, curiosity, and controversy
  - ↪ **Potential backlash channel.** High-profile cases (e.g., *Cleopatra* 2023, *The Little Mermaid* 2023) generated controversy-driven searches in conservative regions

Path forward - Actually Get to the Demand?

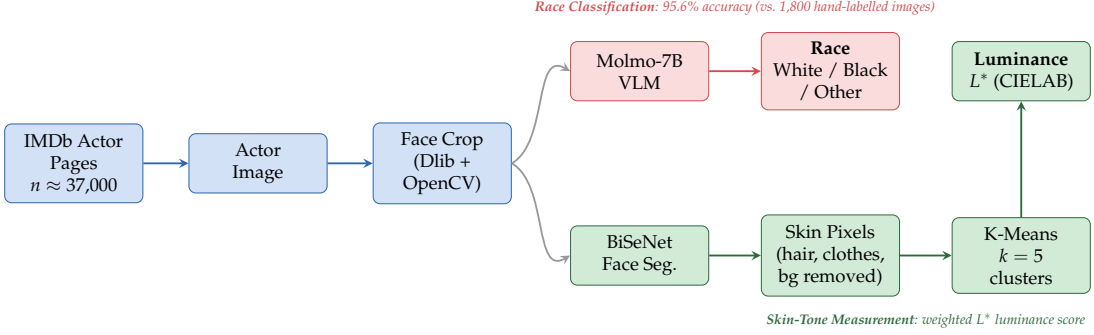
- Trade press (*Variety*, Chronicling America archives) $\rightarrow$ historical critical and box-office reception
- BLS motion picture employment data $\rightarrow$ supply-side composition over time
- Box office data where available $\rightarrow$ closer to realized demand but \$ \$ \$.

From Consumption to Beliefs: A Model Idea

- **What we've shown so far:** Demand (or at least search intensity) for movies correlates with local racial attitudes. But what *happens* after consumption?
- **A sketch:** Viewers don't store the full movie — they compress it into a coarse narrative, subject to limited attention and salience (Bordalo et al. '16, Sims '03). Beliefs then update on this distorted summary, so posteriors inherit bias from both Hollywood's editorial filter and the viewer's own compression.
- **Two predictions worth testing:**
  - ↪ **Heterogeneity in updating.** Viewers with diffuse priors (less direct contact) update more; viewers with concentrated priors are largely unaffected.
  - ↪ **Identity channel.** For in-group viewers, shifted beliefs may affect self-perception and aspirations (Akerlof & Kranton '00).
- **Empirical path:** Testing this likely requires an experiment or panel data on beliefs. We do not have that yet!

Appendix

Data Pipeline (A): Actor Images → Race & Skin Tone



◂ Data Creation

Data Pipeline (A): Skin-Tone Measurement

Original Image

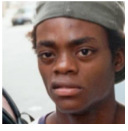
Cropped Face

Skin Mask Overlay

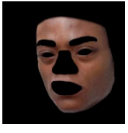
White actor $L^* = 77.02$

Original Image

Cropped Face

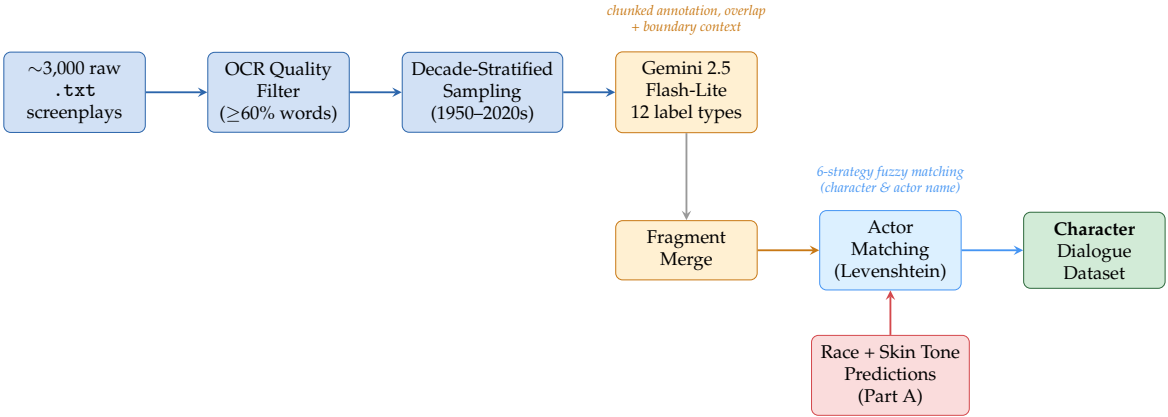


Skin Mask Overlay



Black actor $L^* = 46.18$

Data Pipeline (B): Screenplays → Dialogue Data



◂ Data Creation

Data Pipeline (B): Gemini 2.5 Flash-Lite Prompt

System Prompt

```
You are a screenplay structuring agent.
Classify each non-empty line.  Label lines only ---
never rewrite, paraphrase, or merge them.

ELEMENT TYPES (12)
scene_heading · action · character · character_label ·
parenthetical · dialogue · voice_over · transition · title_card
· metadata · preamble · other

KEY RULES
1.  Return exactly one label per line_id.
2.  speaker ≠ null only for dialogue, voice_over,
parenthetical.
3.  ‘‘DANNY (V.O.)’’ cue → character; lines that follow →
voice_over.
4.  Wrapped lines of the same block → same label.
5.  other only when nothing else fits.
6.  VOICE / V.O. / O.S. cues → always character, never
character_label.
```

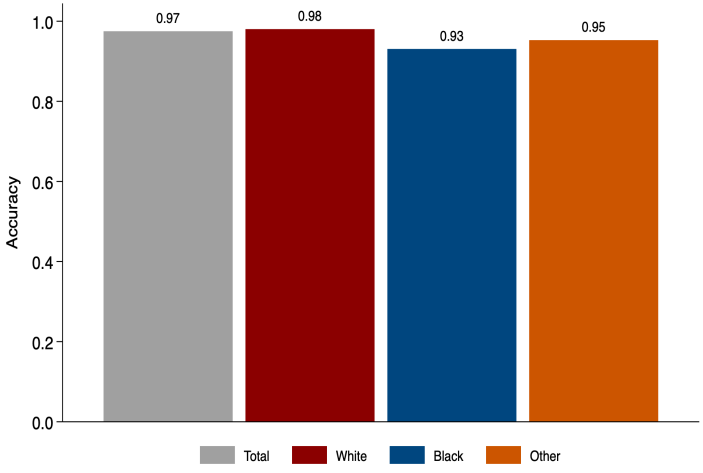
User Prompt (per chunk of ∼30 lines)

```
Script:  American_History_X_0120586
Chunk 3/47

[Boundary context:  last type = ‘action’,
 speaker = None.]

Classify every line.  Return JSON only.

Lines:
42:  ‘‘DEREK’’
43:  ‘‘You have no idea...’’
44:  ‘‘DANNY (V.O.)’’
45:  ‘‘He pulled me back.’’
```

JSON Response

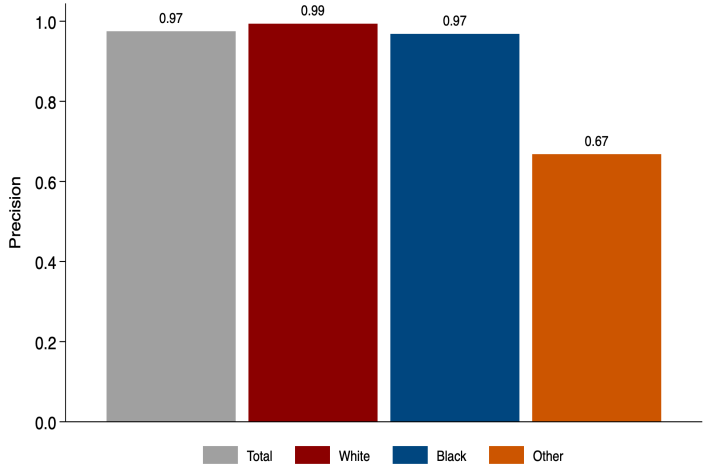
```
{"items":  [
  {"line_id":  42, "type":  "character",
    "speaker":  null},
  {"line_id":  43, "type":  "dialogue",
    "speaker":  "DEREK"},
  {"line_id":  44, "type":  "character",
    "speaker":  null},
  {"line_id":  45, "type":  "voice_over",
    "speaker":  "DANNY"}
]}
```

Accuracy



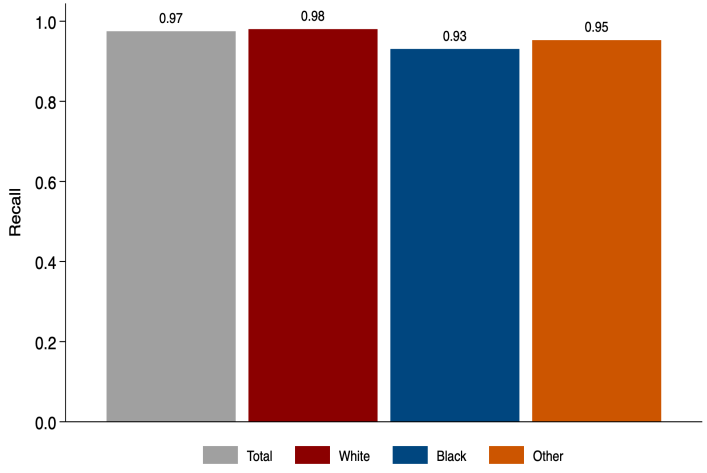
Data Creation

Precision



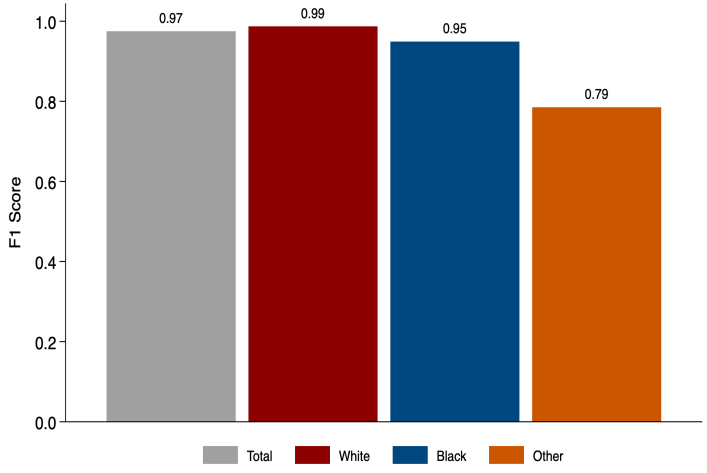
◂ Data Creation

Recall



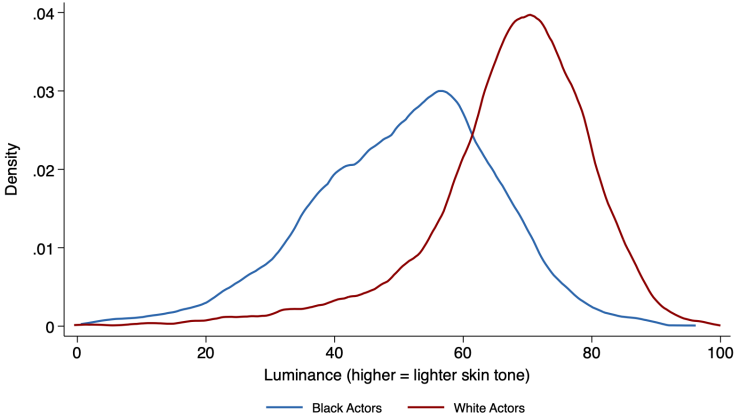
◂ Data Creation

F1 Score



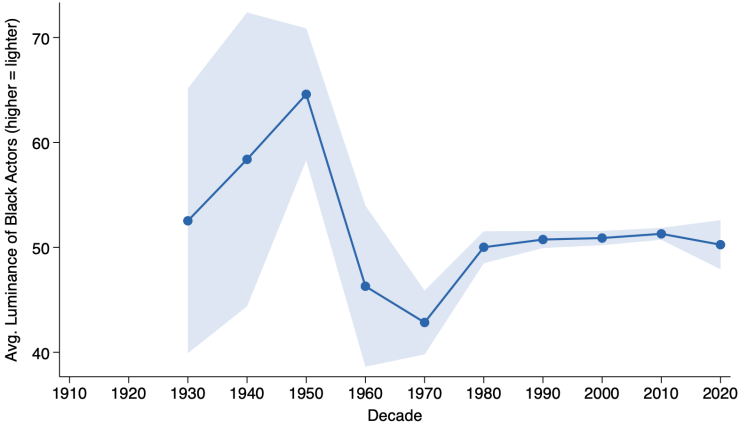
Data Creation

Skin Tone Distribution



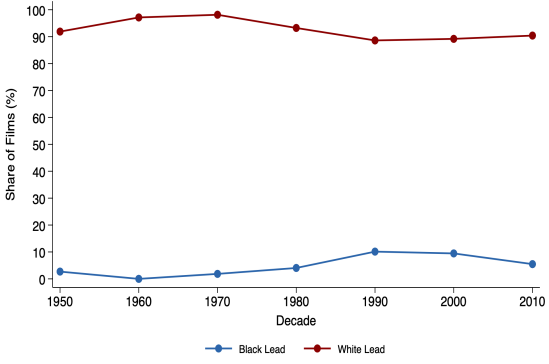
Skin Tone

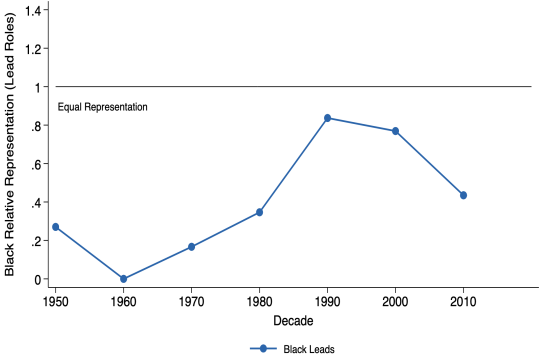
Skin Tone for Blacks Over Time



Skin Tone

Lead Role





◂ Dialogue

Sentiment Analysis: LIWC

- **Linguistic Inquiry and Word Count (LIWC)** is a text analysis tool that measures the psychological and linguistic content of text
- Operates by matching words in a text against curated dictionaries spanning **emotional** (positive/negative affect, anger), **cognitive** (achievement, insight), and **topical** (death, religion, sexuality) dimensions
- For each character, we compute category scores as the **share of dialogue words** belonging to each dimension — e.g. swear words per 100 words spoken

Details

Specification - Justification for log

Proposition 1: Identification of θ under Google Trends normalization Let s_{ij} denote the true search share for movie i in DMA j, and let $I_{ij} = c_i \cdot s_{ij}$ with $c_i \equiv 100/\max_j s_{ij}$ denote the Google Trends index. Consider the estimating equation

$$f(I_{ij}) = \delta_i + \delta_j + \theta\,(R_i \cdot P_j) + \varepsilon_{ij},$$

where δ_i, δ_j are movie and DMA fixed effects.

(i) **Logs.** If the true DGP is $\log s_{ij} = \alpha_i + \alpha_j + \theta\,(R_i \cdot P_j) + u_{ij}$, then the regression with $f(\cdot) = \log(\cdot)$ identifies θ.

(ii) **Levels.** If the true DGP is $s_{ij} = \alpha_i + \alpha_j + \theta\,(R_i \cdot P_j) + u_{ij}$, then the regression with $f(\cdot) = \text{id}$ does *not* identify θ: the coefficient on $R_i \cdot P_j$ is a c_i-weighted average of θ, and the two-way fixed effects fail to absorb a residual term $(c_i - \bar{c})\,\alpha_j$ that is generically correlated with $R_i \cdot P_j$.

Specification - Proof

Part (i). Taking logs of the Trends identity gives $\log I_{ij} = \log c_i + \log s_{ij}$. Substituting the log-linear DGP:

$$\log I_{ij} = \underbrace{\log c_i + \alpha_i}_{\equiv \delta_i} + \underbrace{\alpha_j}_{\equiv \delta_j} + \theta \left(R_i \cdot P_j\right) + u_{ij}.$$

The movie-specific scaling $\log c_i$ enters additively and is absorbed into δ_i. The coefficient on $R_i \cdot P_j$ is exactly θ, identified from within-movie, within-DMA variation in the interaction.

Part (ii). Multiplying the level DGP by c_i:

$$I_{ij} = c_i\alpha_i + c_i\alpha_j + c_i\theta \left(R_i \cdot P_j\right) + c_i u_{ij}.$$

Specification - Proof (continued)

Write $c_i = \bar{c} + (c_i - \bar{c})$ where $\bar{c} \equiv \mathbb{E}[c_i]$. Then

$$I_{ij} = \underbrace{c_i \alpha_i}_{\text{absorbed by } \delta_i} + \underbrace{\bar{c}\,\alpha_j}_{\text{absorbed by } \delta_j} + \bar{c}\,\theta\,(R_i \cdot P_j) + \underbrace{(c_i - \bar{c})\,\alpha_j}_{\text{(A)}} + \underbrace{(c_i - \bar{c})\,\theta\,(R_i \cdot P_j)}_{\text{(B)}} + c_i u_{ij}.$$

Terms (A) and (B) are not absorbed by the two-way fixed effects because $c_i - \bar{c}$ varies at the movie level while α_j varies at the DMA level, so their product varies at the ij level. The probability limit can be derived and shows inconsistency of the estimator:

$$\widehat{\theta}^{\text{OLS}} \xrightarrow{p} \bar{c}\,\theta + \underbrace{\frac{\mathbb{E}[\widetilde{X}_{ij}\,(c_i - \bar{c})\,\alpha_j]}{\mathbb{E}[\widetilde{X}_{ij}^2]}}_{\text{bias}_A} + \underbrace{\theta\,\frac{\mathbb{E}[\widetilde{X}_{ij}^2\,(c_i - \bar{c})]}{\mathbb{E}[\widetilde{X}_{ij}^2]}}_{\text{bias}_B}.$$

◂ Specification

Racial Attitudes Construction: Items and Recoding

Source. CCES 2018 Common Content. All on a 5-point scale (1 = Strongly agree, 5 = Strongly disagree).

Six items combined into a racial progressiveness scale:

→ *Conservative-coded* (agree = conservative):
- 422b: "Racial problems are rare, isolated situations"
- 422e: "Irish, Italians, Jewish overcame prejudice; Blacks should too"
- 422h: "Matter of some not trying hard enough"

→ *Progressive-coded* (agree = progressive):
- 422a: "White people have certain advantages"
- 422f: "Slavery/discrimination created lasting conditions"
- 422g: "Blacks have gotten less than they deserve"

Recoding. Progressive-coded items are reversed ($x \mapsto 6 - x$); conservative-coded items kept as-is. After recoding, every item runs 1–5 with **higher = more racially progressive**.

Racial Attitudes Construction: Aggregation

Step 1: Per-respondent scale. For respondent i, take the mean of recoded items (require ≥ 3 non-missing):

$$S_i = \frac{1}{|A_i|} \sum_{k \in A_i} \widetilde{x}_{ik} \in [1, 5]$$

where A_i is the set of items respondent i answered and $\widetilde{x}_{ik}$ is the recoded response.

Step 2: DMA-level aggregation. For DMA j, take the weighted mean across respondents using CCES survey weights w_i:

$$P_j = \frac{\sum_{i \in j} w_i \, S_i}{\sum_{i \in j} w_i}$$

◂ Specification

Identifying Movies in Google Trends

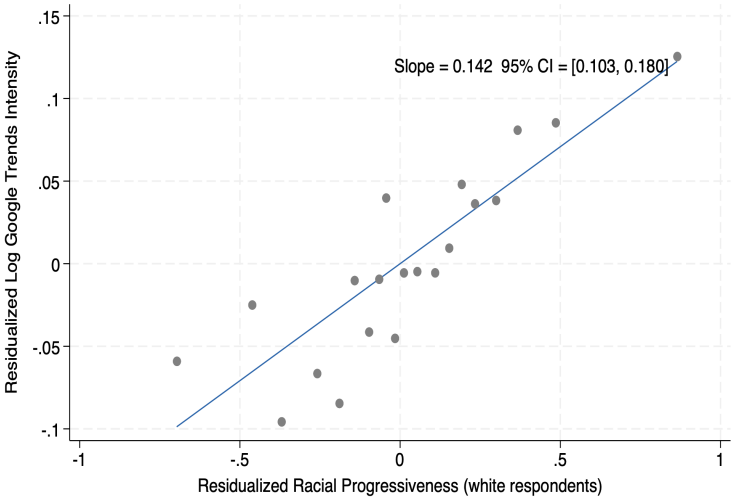
Problem. Title-based search is contaminated by generic usage for many movies in our sample (e.g., *Her*, *Up*, *Gravity*, *Frozen*, *Parasite*, *1917*).

Our approach. We map each IMDb ID to its Google Knowledge Graph topic ID, which Google Trends accepts directly as a disambiguated query:

1. Query the Knowledge Graph Search API with `"{title} {year} movie"`, restricted to entities typed as `Movie` → recover Knowledge Graph ID.
2. Select the top-ranked result and record its topic ID (format `/m/xxxxx` or `/g/xxxxx`).
3. Validate by comparing the returned entity name to our input title (normalized). Flag cases below a 0.75 similarity threshold for manual review.

◂ Google Trends

Search interest and racial attitudes



◂ Preliminary Evidence