rACE Coding Scheme for Open-Ended Responses to Race-Related Questions

Francy Luna Diaz

June 2026

Coding Framework: The rACE Coding Scheme

Construct Definition

To measure open-ended racial discourse, I created the *rACE coding scheme*. The name highlights that *racial* is the common modifier across the three dimensions: racial **Attribution**, racial **Color-blindness**, and racial **Explicitness**. The scheme is informed by prior work on color-blind racism (Bonilla-Silva, 2017), color-blind racial ideology (Neville et al., 2013), and implicit racial appeals (Mendelberg, 2001). However, rACE is designed as a new text-coding framework for open-ended survey responses rather than as a reproduction of any single existing scale.

rACE assigns each open-ended response three analytically distinct scores. The *racial Attribution Score* captures whether the response explains racial inequality through structural causes or through individual or group-level causes. The *racial Color-blindness Score* captures whether the response is color-conscious or color-blind in its interpretation of race and racism. The *racial Explicitness Score* captures whether racial meaning is implicit, mixed, or explicit.

The first two scores locate each response on a Cartesian plane: the racial Color-blindness Score is plotted on the X-axis and the racial Attribution Score is plotted on the Y-axis. The racial Explicitness Score is used to visualize the form of racial expression through a color gradient when it is available. Responses that can be assigned racial Attribution and racial Color-blindness scores but cannot be assigned a racial Explicitness score can still be plotted on the Cartesian plane using a distinct marker for missing explicitness. Respondent characteristics, such as partisanship, can be added as additional visual encodings, such as point shape, but are not part of the rACE scores themselves.

The coding scheme draws on several related theoretical concepts. Bonilla-Silva's theory of color-blind racism identifies central frames such as abstract liberalism, naturalization, cultural racism, and minimization of racism. Neville and colleagues distinguish between color-evasion, or the denial of racial difference through appeals to sameness, and power-evasion, or the denial of racism by emphasizing equal opportunity while ignoring institutional power. Mendelberg's theory of racial appeals distinguishes between explicit racial messages and implicit or coded racial messages, where racial meaning is communicated through ostensibly nonracial language.

The purpose of the rACE coding scheme is not to determine whether a respondent is personally racist. Instead, the goal is to classify the content, attributional logic, racial interpretation, and expressive form of each open-ended response. This multidimensional approach is useful for capturing the nuance of race-related discourse in the United States. Contemporary racial discourse takes multiple forms: it may appear as explicit racial hostility, implicit or coded racial appeals, color-blind minimization, structural recognition, individual-blame attribution, or claims of reverse discrimination. A single linear scale would risk collapsing these distinct manifestations into one measure and obscuring meaningful variation across responses. By separating attributional logic, racial interpretation, and explicitness, the rACE coding scheme allows the analysis to capture both the substance of racial reasoning and the form through which it is expressed.

1 rACE Coding Structure

1.1 Overview

Each response receives three separate rACE scores:

$$A_i = \text{racial Attribution}$$

$$C_i = \text{racial Color-blindness}$$

$$E_i = \text{racial Explicitness}$$

Each rACE dimension is conceptually continuous, but responses are coded using a five-point ordinal approximation to reduce false precision and improve interpretability:

$$A_i, C_i, E_i \in \{-1, -0.5, 0, 0.5, 1\}$$

The coding procedure also includes two response-quality flags:

$$U_i = \text{Unclear Response Flag}$$

$$P_i = \text{Partial Coding Flag}$$

where:

$$U_i, P_i \in \{0, 1\}$$

The unclear-response flag identifies responses that are fully unclear, irrelevant, blank, nonsensical, or impossible to classify. The partial coding flag identifies responses where at least one rACE dimension can be coded but one or more other dimensions cannot be coded. These flags are not part of the rACE scores themselves. They are response-quality indicators used to distinguish fully codable, partially codable, and fully non-substantive responses.

Table 1: rACE Dimensions and Response-Quality Flags

Measure	Var	Values	Interpretation
racial Attribution	A_i	$-1, -0.5, 0, 0.5, 1$	Five-point score ranging from structural attribution (-1) to individual or group-blame attribution (1).
racial Color-blindness	C_i	$-1, -0.5, 0, 0.5, 1$	Five-point score ranging from color-conscious reasoning (-1) to color-blind reasoning (1).
racial Explicitness	E_i	$-1, -0.5, 0, 0.5, 1$	Five-point score ranging from implicit or coded racial meaning (-1) to explicit racial meaning (1).
Unclear Response Flag	U_i	$0, 1$	Binary indicator where 1 identifies responses that are fully unclear, irrelevant, blank, nonsensical, or impossible to classify.
Partial Coding Flag	P_i	$0, 1$	Binary indicator where 1 identifies responses for which at least one rACE dimension can be coded but one or more other dimensions must be assigned missing values.

The rACE dimensions should generally be analyzed separately. Explicitness should not be treated as equivalent to color-blindness. A response can be explicit and color-conscious, explicit and color-blind, implicit and color-blind, or implicit and structurally oriented.

2 Dimension A: racial Attribution

2.1 Definition

The racial Attribution dimension measures whether the response explains racial inequality through structural causes or through individual or group-level causes.

$$A_i \in \{-1, -0.5, 0, 0.5, 1\}$$

where:

$$-1 = \text{strong structural attribution}$$

$$-0.5 = \text{mostly structural attribution}$$

0 = mixed, balanced, or ambiguous attribution

0.5 = mostly individual or group-blame attribution

1 = strong individual or group-blame attribution

Table 2: rACE-A: racial Attribution

Score	**Label**	**Definition**
-1	Strong structural attribution	The response clearly and primarily explains racial inequality through institutions, policies, history, discrimination, segregation, unequal schools, housing, labor markets, policing, wealth gaps, or other structural conditions.
-0.5	Mostly structural attribution	The response mostly explains racial inequality through structural causes, but the structural explanation is less developed, more moderate, or contains minor non-structural elements.
0	Mixed, balanced, or ambiguous attribution	The response contains both structural and individual or group-level explanations, or does not clearly identify the cause of racial inequality.
0.5	Mostly individual or group-blame attribution	The response mostly explains racial inequality through individual effort, choices, values, culture, family structure, work ethic, motivation, criminality, dependency, irresponsibility, or group-level deficiency, but may contain some limited acknowledgment of structural factors.
1	Strong individual or group-blame attribution	The response clearly and primarily explains racial inequality through individual behavior, culture, family structure, effort, values, choices, or group-level deficiency, with little or no acknowledgment of structural causes.

Examples:

$A = -1$**:** "Racial disparities are caused by unequal schools, housing discrimination, and the long-term effects of segregation."

$A = -0.5$**:** "Discrimination and unequal schools still make it harder for some groups to get ahead."

$A = 0$**:** "Racism still exists, but people also need to make better choices and work hard."

$A = 0.5$**:** "There are some barriers, but a lot of the problem is that people do not take education seriously."

$A = 1$**:** "The problem is not racism. People just need to value education and take responsibility."

3 Dimension C: racial Color-blindness

3.1 Definition

The racial Color-blindness dimension measures whether the response is color-conscious or color-blind in its interpretation of race and racism.

$$C_i \in \{-1, -0.5, 0, 0.5, 1\}$$

where:

$-1 =$ strong color-conscious reasoning

$-0.5 =$ mostly color-conscious reasoning

$0 =$ mixed, balanced, or ambiguous reasoning

$0.5 =$ mostly color-blind reasoning

$1 =$ strong color-blind reasoning

Table 3: rACE-C: racial Color-blindness

Score	Label	Definition
-1	Strong color-conscious reasoning	The response clearly recognizes race as socially, politically, or institutionally consequential. It may acknowledge racism, racial inequality, racialized experience, unequal treatment, racial privilege, or the value of race-conscious remedies.
-0.5	Mostly color-conscious reasoning	The response acknowledges that race or racism matters, but does so in a more limited, vague, or moderate way.
0	Mixed, balanced, or ambiguous reasoning	The response contains both color-conscious and color-blind elements, is too vague to classify, or does not clearly indicate whether race is treated as consequential.
0.5	Mostly color-blind reasoning	The response tends to minimize, avoid, or downplay the relevance of race, but does not fully deny the significance of race or racism.
1	Strong color-blind reasoning	The response clearly denies, avoids, minimizes, naturalizes, or rationalizes the relevance of race. It may emphasize sameness, universal humanity, "not seeing race," formal equality, reverse racism, or the claim that talking about race is divisive or unnecessary.

Examples:

$C = -1$**:** "Race still shapes people's opportunities because institutions have treated groups differently."

$C = -0.5$**:** "Race still matters in some areas, even if things have improved."

$C = 0$**:** "Race can matter sometimes, but I think other things matter too."

$C = 0.5$**:** "Racism exists, but people focus too much on race now."

$C = 1$**:** "I do not see race. We are all just people, and talking about race only divides us."

4 Dimension E: racial Explicitness

4.1 Definition

The racial Explicitness dimension measures whether the racial meaning of the response is implicit, mixed, or explicit.

$$E_i \in \{-1, -0.5, 0, 0.5, 1\}$$

where:

$$-1 = \text{fully implicit or coded racial meaning}$$

$$-0.5 = \text{mostly implicit or coded racial meaning}$$

$$0 = \text{mixed or partly explicit racial meaning}$$

$$0.5 = \text{mostly explicit racial meaning}$$

$$1 = \text{fully explicit racial meaning}$$

Unlike racial Attribution and racial Color-blindness, racial Explicitness is not a measure of whether the response is more or less color-blind. It captures the form of racial expression. A response may be explicitly color-conscious, explicitly color-blind, implicitly color-blind, or implicit and structurally oriented.

Table 4: rACE-E: racial Explicitness

Score	Label	Definition
-1	Fully implicit or coded	The response does not explicitly name race, racism, racial groups, discrimination, diversity, or race-conscious policy, but communicates racial meaning through coded or racially associated language.
-0.5	Mostly implicit or coded	The response is mostly indirect or coded but contains some contextual cues that make racial meaning partially visible.
0	Mixed or partly explicit	The response contains some direct racial language but also relies on coded, indirect, or ambiguous language.
0.5	Mostly explicit	The response directly names race, racism, racial groups, discrimination, racial inequality, diversity, affirmative action, White people, Black people, Latinos, Asians, or race-conscious policy, but also contains some indirect or ambiguous racial meaning.
1	Fully explicit	The response directly and clearly names race, racism, racial groups, racial discrimination, racial inequality, racial privilege, diversity, affirmative action, or race-conscious policy.

Examples:

$E = -1$**:** "People in those communities need to stop depending on handouts and take responsibility."

$E = -0.5$**:** "Some communities have problems with culture and responsibility."

$E = 0$**:** "Discrimination can happen, but some communities also need to take more responsibility."

$E = 0.5$**:** "Discrimination against Black people exists, but some communities also need to take responsibility."

$E = 1$**:** "Black Americans still face discrimination in housing and policing."

5 Unclear Response Flag

In addition to the three rACE dimensions, the LLM should assign an unclear-response flag. This flag identifies responses that cannot be substantively classified at all.

$$U_i \in \{0, 1\}$$

where:

$$0 = \text{substantively codable or partially codable response}$$

$$1 = \text{fully unclear, irrelevant, blank, nonsensical, or impossible to classify}$$

The unclear-response flag is not part of the rACE scores. Instead, it is a response-quality indicator used to identify cases that should be reviewed, excluded, or analyzed separately.

A response should receive $U = 1$ when it is:

- blank or nearly blank;
- irrelevant to the prompt;
- nonsensical or impossible to interpret;
- sarcastic in a way that cannot be decoded;
- too vague to determine any racial Attribution, racial Color-blindness, or racial Explicitness;
- a refusal or non-answer, such as "I don't know," "no opinion," or "not sure."

When $U = 1$, assign missing values for the rACE dimensions:

$$A = \text{NA}, \quad C = \text{NA}, \quad E = \text{NA}$$

and set confidence to `low`. These cases should not be interpreted as substantively mixed responses. They indicate that the response could not be classified.

6 Partial Coding Flag

Some responses may contain enough information to code one or two rACE dimensions, but not all three. These responses should not be treated as fully unclear if at least one dimension can be substantively coded. Instead, the LLM should assign available dimension scores and use missing values for dimensions that cannot be classified.

$$P_i \in \{0, 1\}$$

where:

$$0 = \text{all applicable rACE dimensions are codable}$$

$$1 = \text{partially codable response}$$

A response should receive $P = 1$ when at least one rACE dimension can be coded but one or more other dimensions cannot be coded. For dimensions that cannot be classified, assign missing values as needed:

$$A = \text{NA}, \quad C = \text{NA}, \quad \text{or} \quad E = \text{NA}$$

The partial coding flag is distinct from the unclear-response flag. Use $U = 1$ only when the entire response is fully unclear, irrelevant, blank, nonsensical, or impossible to classify. Use $P = 1$ when the response is partially interpretable but does not provide enough information to score all three rACE dimensions.

For example, a response such as "Black people" may receive:

$$A = \text{NA}, \quad C = \text{NA}, \quad E = 1, \quad U = 0, \quad P = 1$$

because the response explicitly names a racial group but does not provide enough content to classify racial Attribution or racial Color-blindness.

7 Visualizing rACE Scores

The racial Attribution and racial Color-blindness dimensions can be visualized as a Cartesian plane. The racial Attribution score, A_i, is plotted on the Y-axis, and the racial Color-blindness score, C_i, is plotted on the X-axis. The racial Explicitness score, E_i, can be displayed using a grayscale gradient when it is available.

$$x_i = C_i$$

$$y_i = A_i$$

$$\text{color intensity}_i = E_i$$

Respondent-level characteristics can be added as additional visual encodings. For example, partisanship can be represented through point shape, allowing the visualization to show whether responses with similar rACE coordinates cluster by party identification. These respondent characteristics are not part of the rACE scores themselves.

Responses where $U_i = 1$ should generally be excluded from the main rACE visualization or displayed separately as non-substantive responses. Responses where $P_i = 1$ can still be plotted when the dimensions needed for the specific visualization are available. For the main Cartesian visualization, a response can be plotted as long as both A_i and C_i are available, because these two scores determine the response's location on the plane.

If E_i is missing but A_i and C_i are available, the response should still be plotted on the Cartesian plane. In this case, the point should use a distinct visual marker for missing explicitness, such as a hollow point. This allows the visualization to retain substantively codable responses without misrepresenting the missing E_i value as implicit, mixed, or explicit.

If either A_i or C_i is missing, the response cannot be located on the main Cartesian plane and should be omitted from that visualization or displayed in a separate summary of partially codable responses. However, the available dimension or dimensions can still be analyzed descriptively.

7.1 Interpreting the Cartesian Plane

The racial Attribution and racial Color-blindness dimensions generate four main regions, shown in Figure 1.

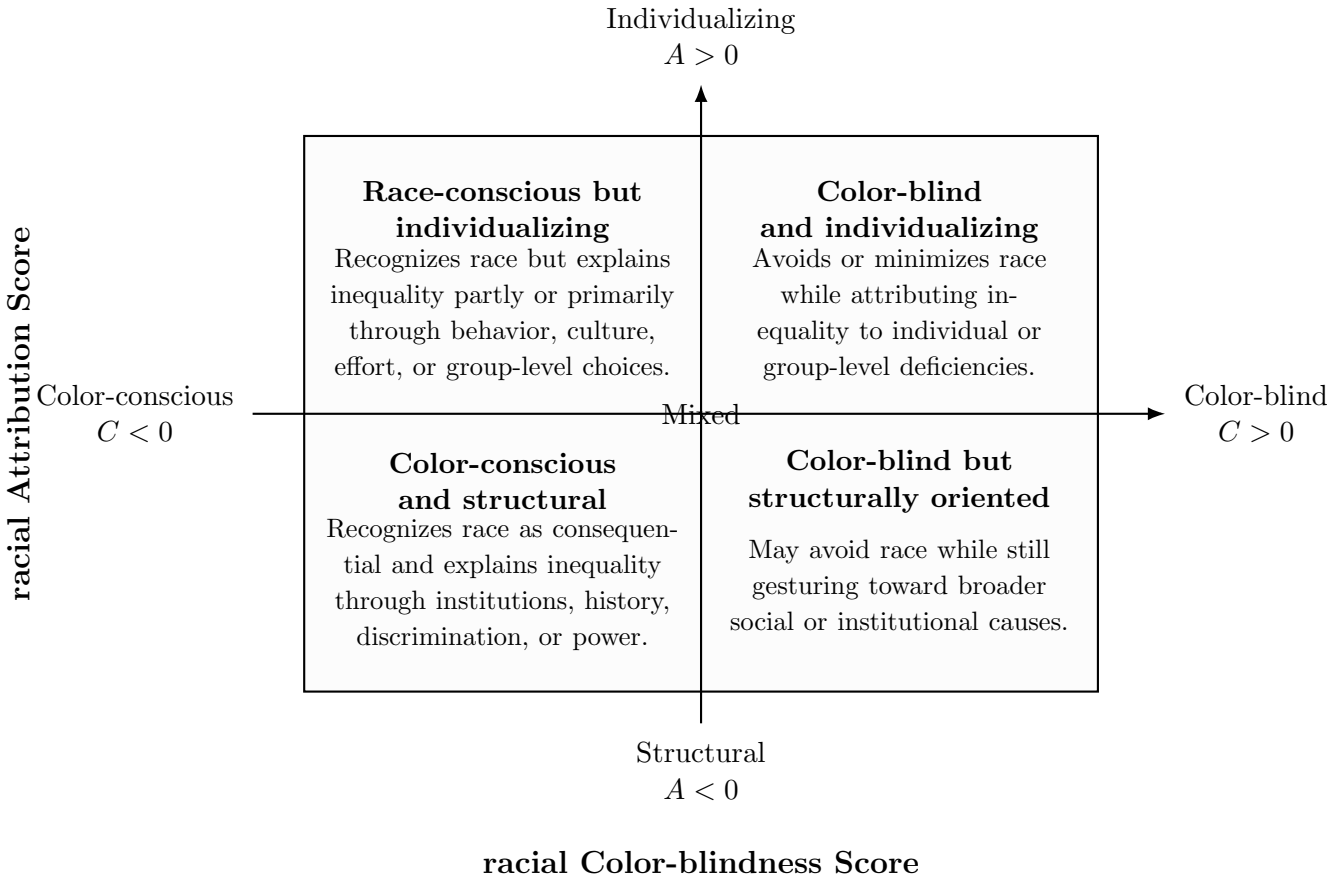

Figure 1: Interpretation of rACE Cartesian Regions

8 Relationship to Existing Theoretical Dimensions

The rACE coding scheme incorporates several concepts from the literature.

8.1 Color-Evasion

Color-evasion appears primarily in the racial Color-blindness dimension. Responses are more color-blind when they deny or avoid the relevance of race by emphasizing sameness, universal humanity, or the claim that race should not matter.

8.2 Power-Evasion

Power-evasion appears in both the racial Color-blindness and racial Attribution dimensions. Responses are more color-blind and individualizing when they deny institutional racism, racial privilege, or unequal access to resources while presenting society as basically fair.

8.3 Abstract Liberalism

Abstract liberalism appears when respondents use principles such as equal opportunity, individualism, choice, meritocracy, or limited government intervention in an abstract way that ignores racial inequality. Abstract liberalism usually pushes the racial Color-blindness score toward color-blindness and, when tied to merit or individual effort, pushes the racial Attribution score toward individual attribution.

8.4 Naturalization

Naturalization appears when respondents treat racial separation, inequality, or homogeneity as natural, inevitable, or merely the result of normal human preferences. Naturalization usually pushes the racial Color-blindness score toward color-blindness because it removes racism, institutions, history, and power from the explanation.

8.5 Cultural Racism

Cultural racism appears when respondents explain racial inequality through alleged group-level cultural deficiencies, values, behaviors, family structures, or work ethic. Cultural racism usually pushes the racial Attribution score toward individual or group-blame attribution. If the response also denies racism or minimizes race, it should push the racial Color-blindness score toward color-blindness.

8.6 Minimization of Racism

Minimization appears when respondents acknowledge racism only to reduce its contemporary importance or deny its relevance to present-day inequality. Minimization usually pushes the racial Color-blindness score toward color-blindness. If the response replaces racism with effort, culture, or personal responsibility, it also pushes the racial Attribution score toward individual attribution.

8.7 Reverse Racism and White Victimhood

Reverse-racism and White-victimhood claims appear when respondents argue that White people are now the primary victims of racism or that racial justice policies unfairly disadvantage White people. These claims are treated as a form of racial color-blindness because they deny, minimize, or invert racial power relations.

Examples:

"Affirmative action is reverse racism."
"White people are discriminated against now."
"Diversity programs are unfair to White people."
"Minorities get special treatment."
"The real racism today is against Whites."

These responses should generally receive a high racial Color-blindness score because they reject or invert the relevance of institutional racism and racial privilege. They should also receive a high racial Explicitness score when they directly name race, racial groups, affirmative action, diversity, or race-conscious policy. The racial Attribution score should depend on whether the response explains racial inequality through structural causes, mixed causes, or individual/group-blame causes.

8.8 Subversion of Racism Claims

Subversion appears when respondents treat naming racism as itself racist, divisive, or illegitimate. Subversion usually pushes the racial Color-blindness score toward color-blindness because it delegitimizes discussion of racism.

8.9 Implicit or Coded Racial Messaging

Implicit or coded racial messaging appears when respondents communicate racial meaning without direct racial language. This dimension is captured by the racial Explicitness score. Do not code a term as racialized merely because it appears in the response. Code it as implicit or coded racial messaging only when the surrounding context links the term to racial inequality, racial minorities, racialized policy, or racial resentment.

9 Decision Rules

The LLM should apply the following decision rules:

1. Code the full meaning of the response, not isolated words.
2. Do not infer the respondent's intent beyond the text.
3. Code racial Attribution, racial Color-blindness, and racial Explicitness separately. A response can be color-conscious but individualizing, or color-blind but not clearly individualizing.

4. Code racial Explicitness separately from substance. Explicit racial content is not automatically color-blind. For example, "Black Americans still face discrimination" is explicit but color-conscious.

5. Universalist statements such as "everyone is equal" or "I treat everyone the same" should push the racial Color-blindness score toward color-blindness only when they avoid, deny, or dismiss racial inequality.

6. Opposition to a specific race-conscious policy should not automatically be coded as color-blind. Code it as color-blind only when the opposition is justified through abstract liberalism, meritocracy, reverse racism, minimization of racism, or denial of institutional inequality.

7. Claims of "reverse racism," White victimhood, unfair anti-White discrimination, or "special treatment" for racial minorities should generally push the racial Color-blindness score toward $C = 0.5$ or $C = 1$, because they deny or invert racial power relations. If these claims explicitly name race, racial groups, affirmative action, diversity policy, or race-conscious remedies, they should also push the racial Explicitness score toward $E = 0.5$ or $E = 1$.

8. If a response recognizes racism but primarily blames racial minorities for inequality, assign a more color-conscious score on racial Color-blindness but a more individualizing score on racial Attribution.

9. If a response acknowledges racism only to minimize it, such as "racism exists, but it is not really the issue anymore," push the racial Color-blindness score toward color-blindness.

10. If a response denies institutional racism or White privilege, push the racial Color-blindness score toward color-blindness.

11. If a response uses explicit racial stereotypes, biological essentialism, or group-deficiency claims, push the racial Attribution score toward individual or group-blame attribution and set the racial Explicitness score closer to explicit.

12. If the response is blank, irrelevant, nonsensical, sarcastic but unclear, or impossible to interpret, assign NA for A, C, and E, set $U = 1$, set $P = 0$, set `quadrant = unclear/non-substantive`, and mark confidence as low. These cases should not be interpreted as substantively mixed responses.

13. If at least one rACE dimension can be coded but one or more dimensions cannot be coded, assign the available dimension scores, assign NA for dimensions that cannot be classified, set $U = 0$, set $P = 1$, and mark confidence as low or medium depending on the amount of interpretable content.

14. If A and C are both available but E is missing, the response can still be assigned to the appropriate Cartesian region and plotted on the main rACE plane. The missing E value should be represented visually using a hollow point rather than a grayscale shade.

10 LLM Coding Prompt

The following prompt should be given to each initial LLM coder. The same prompt should be used for all initial models, changing only the model name indicated in the output requirements. The LLM should receive the blinded input file containing only `caseid` and `response_text`. The full protocol, respondent covariates, treatment assignment, partisanship, race, weights, and other respondent characteristics should not be provided to the initial coding models.

> You are coding open-ended survey responses using the rACE coding scheme.
>
> You will receive a CSV file with two columns:
>
> ```
> caseid
> response_text
> ```
>
> Your task is to code each response independently. Do not use any information outside the response text. Do not infer the respondent's partisanship, ideology, race, education, treatment condition, or demographic characteristics. Code only the content of the open-ended response.
>
> For this coding task, code only two rACE dimensions: racial Attribution and racial Color-blindness. Do not code racial Explicitness.
>
> **1. racial Attribution**
>
> This dimension measures whether the response explains racial inequality through structural causes or through individual or group-level causes.
>
> Allowed values:
>
> ```
> -1   = strong structural explanation
> -0.5 = mostly structural explanation
> 0    = mixed, balanced, or ambiguous attribution
> 0.5  = mostly individual or group-blame explanation
> 1    = strong individual or group-blame explanation
> null = this dimension cannot be coded from the response
> ```
>
> Assign lower values when the response explains racial inequality through institutions, policies, history, discrimination, segregation, unequal schools, housing, labor markets, policing, wealth gaps, or other structural conditions.
>
> Assign higher values when the response explains racial inequality through individual effort, choices, values, culture, family structure, work ethic, motivation, criminality, dependency, irresponsibility, or group-level deficiency.
>
> **2. racial Color-blindness**
>
> This dimension measures whether the response is color-conscious or color-blind in its interpretation of race and racism.
>
> Allowed values:

```
-1   = strong color-conscious reasoning
-0.5 = mostly color-conscious reasoning
0    = mixed, balanced, or ambiguous reasoning
0.5  = mostly color-blind reasoning
1    = strong color-blind reasoning
null = this dimension cannot be coded from the response
```

Assign lower values when the response recognizes race as socially, politically, historically, or institutionally consequential. This includes recognition of racism, racial inequality, racialized experience, unequal treatment, racial privilege, or the value of race-conscious remedies.

Assign higher values when the response denies, avoids, minimizes, naturalizes, or rationalizes the relevance of race or racism. This includes "not seeing race," claims that talking about race is divisive, claims that racism is mostly in the past, denial of institutional racism, reverse-racism claims, White-victimhood claims, or claims that race-conscious policies unfairly advantage racial minorities.

Response-Quality Flags

Use `unclear_flag = 1` only when the entire response is non-substantive or impossible to classify. Examples include blank responses, irrelevant responses, nonsensical responses, undecodable sarcasm, or answers such as "don't know," "not sure," "no opinion," "none," or "n/a."

When `unclear_flag = 1`, assign `null` for `race_attribution` and `race_colorblindness`; set `partial_coding_flag = 0`; set `quadrant = unclear/non-substantive`; and set `confidence = low`.

Use `partial_coding_flag = 1` when one dimension can be coded but the other dimension cannot be coded. Assign the available dimension score and assign `null` for the dimension that cannot be coded. If either `race_attribution` or `race_colorblindness` is `null`, set `quadrant = not plotted`.

Do not use 0 when a dimension is absent. Use `null` when the response does not contain enough information to score that dimension. Use 0 only when the response contains mixed, balanced, or ambiguous information relevant to that dimension.

Quadrant Rules

Use one of the following values for `quadrant`:

```
color-conscious structural
color-blind structural
race-conscious individualizing
color-blind individualizing
mixed/unclear
unclear/non-substantive
not plotted
```

Assign quadrant using `race_colorblindness` and `race_attribution`:

```
If race_colorblindness < 0 and race_attribution < 0:
quadrant = color-conscious structural

If race_colorblindness > 0 and race_attribution < 0:
quadrant = color-blind structural

If race_colorblindness < 0 and race_attribution > 0:
quadrant = race-conscious individualizing

If race_colorblindness > 0 and race_attribution > 0:
quadrant = color-blind individualizing

If race_colorblindness = 0 or race_attribution = 0:
quadrant = mixed/unclear

If unclear_flag = 1:
quadrant = unclear/non-substantive

If race_colorblindness is null or race_attribution is null, and unclear_flag is not
quadrant = not plotted
```

Theoretical Dimensions

For `primary_dimension`, choose one of the following:

```
color-evasion
power-evasion
abstract liberalism
naturalization
cultural racism
minimization of racism
reverse racism/White victimhood
subversion
structural recognition
none
```

Use `secondary_dimensions` for any additional dimensions present. Separate multiple values with semicolons.

Output Requirements

Return a CSV only. Do not return markdown. Do not return explanations outside the CSV. Do not summarize the responses. Do not skip any rows. Return one row for every row in the uploaded CSV file.

The output CSV must contain exactly these columns:

```
caseid
response_text
model_name
race_attribution
race_colorblindness
unclear_flag
partial_coding_flag
quadrant
confidence
primary_dimension
secondary_dimensions
rationale
flags
```

Use the model name: `[MODEL NAME]`

Allowed values for `confidence`:

```
high
medium
low
```

The `rationale` should be one short sentence.

The `flags` column can include values such as:

```
ambiguous
sarcasm
non-substantive
fully unclear
partial coding
policy-only
explicit stereotype
reverse racism
White victimhood
coded language
```

Separate multiple flags with colons. Separate columbs with semicolobs. Leave blank if no flags apply.

Now code every row in the uploaded CSV file.

11 Output Variable Definitions

The LLM output should be returned as a CSV file with one row per response. The CSV should preserve the original `caseid` and `response_text` columns and add the coding variables described below.

- `model_name`: The name of the LLM coder that produced the row.
- `race_attribution`: A numeric score taking values -1, -0.5, 0, 0.5, or 1. Use `null` when this dimension cannot be coded.
- `race_colorblindness`: A numeric score taking values -1, -0.5, 0, 0.5, or 1. Use `null` when this dimension cannot be coded.
- `race_explicitness`: A numeric score taking values -1, -0.5, 0, 0.5, or 1. Use `null` when this dimension cannot be coded. If this value is `null` but `race_attribution` and `race_colorblindness` are available, the response can still be plotted on the Cartesian plane using a distinct marker for missing explicitness.
- `unclear_flag`: A binary indicator taking values 0 or 1, where 1 indicates a fully unclear, irrelevant, blank, nonsensical, or impossible-to-classify response.
- `partial_coding_flag`: A binary indicator taking values 0 or 1, where 1 indicates that at least one rACE dimension can be coded but one or more other dimensions cannot be coded.
- `quadrant`: One of `color-conscious structural`, `color-blind structural`, `race-conscious individualizing`, `color-blind individualizing`, `mixed/unclear`, `unclear/non-substantive`, or `not plotted`.
- `confidence`: One of `high`, `medium`, or `low`.
- `primary_dimension`: The main theoretical feature driving the classification. Use one of: `color-evasion`, `power-evasion`, `abstract liberalism`, `naturalization`, `cultural racism`, `minimization of racism`, `reverse racism/White victimhood`, `subversion`, `implicit/coded racial messaging`, `structural recognition`, `explicit racial reference`, or `none`.
- `secondary_dimensions`: Any additional theoretical dimensions present in the response. If multiple dimensions are present, separate them with semicolons.
- `rationale`: A one-sentence explanation of the classification.
- `flags`: Any relevant flags, such as `ambiguous`, `sarcasm`, `non-substantive`, `fully unclear`, `partial coding`, `missing explicitness`, `policy-only`, `explicit stereotype`, `reverse racism`, `White victimhood`, or `coded language`. If multiple flags are present, separate them with semicolons.

12 Aggregation Across Multiple LLM Coders

Each response should be independently coded by four LLMs using identical coding instructions. The final score for each rACE dimension should be assigned using the median score across models, excluding missing values when at least two coders provide valid scores.

Let A_{i1}, A_{i2}, A_{i3}, and A_{i4} represent the racial Attribution scores assigned to response i by four independent LLM coders. The final racial Attribution score is:

$$A_i^{median} = \text{median}(A_{i1}, A_{i2}, A_{i3}, A_{i4})$$

computed over non-missing values when possible. Similarly:

$$C_i^{median} = \text{median}(C_{i1}, C_{i2}, C_{i3}, C_{i4})$$

$$E_i^{median} = \text{median}(E_{i1}, E_{i2}, E_{i3}, E_{i4})$$

If fewer than two coders provide a valid score for a given dimension, the final value for that dimension should remain missing unless adjudicated.

For the unclear-response flag, the final value can be assigned using majority rule:

$$U_i^{final} = \begin{cases} 1 & \text{if at least three of four coders assign } U = 1 \\ 0 & \text{otherwise} \end{cases}$$

For the partial coding flag, the final value can also be assigned using majority rule:

$$P_i^{final} = \begin{cases} 1 & \text{if at least three of four coders assign } P = 1 \\ 0 & \text{otherwise} \end{cases}$$

Cases of substantial inter-model disagreement should be sent to the adjudication model. For adjudicated cases, the adjudicator's final scores replace the median model scores.

13 Disagreement Flag

A response should be flagged for adjudication when coders disagree substantially on racial Attribution, racial Color-blindness, or racial Explicitness. For each dimension, disagreement is calculated as the range across the four initial LLM coders, excluding missing values.

$$\text{Range}_{A_i} = \max(A_{i1}, A_{i2}, A_{i3}, A_{i4}) - \min(A_{i1}, A_{i2}, A_{i3}, A_{i4})$$

$$\text{Range}_{C_i} = \max(C_{i1}, C_{i2}, C_{i3}, C_{i4}) - \min(C_{i1}, C_{i2}, C_{i3}, C_{i4})$$

$$\text{Range}_{E_i} = \max(E_{i1}, E_{i2}, E_{i3}, E_{i4}) - \min(E_{i1}, E_{i2}, E_{i3}, E_{i4})$$

A response should be flagged for adjudication when:

$$\text{Range}_{A_i} \geq 1$$

or:

$$\text{Range}_{C_i} \geq 1$$

or:

$$\text{Range}_{E_i} \geq 1$$

A response should also be flagged if the coders disagree about whether the response is fully unclear or partially codable:

$$\max(U_{i1}, U_{i2}, U_{i3}, U_{i4}) - \min(U_{i1}, U_{i2}, U_{i3}, U_{i4}) = 1$$

or:

$$\max(P_{i1}, P_{i2}, P_{i3}, P_{i4}) - \min(P_{i1}, P_{i2}, P_{i3}, P_{i4}) = 1$$

Flagged cases should be sent to the adjudicator model using the same codebook. The adjudicator should assign final racial Attribution, racial Color-blindness, racial Explicitness, unclear-response flag, and partial coding flag values by applying the codebook directly, not by averaging the prior model responses.

14 Disagreement Flag

A response should be flagged for adjudication when coders disagree substantially on racial Attribution, racial Color-blindness, or racial Explicitness. For each dimension, disagreement is calculated as the range across the four initial LLM coders, excluding missing values.

$$\text{Range}_{A_i} = \max(A_{i1}, A_{i2}, A_{i3}, A_{i4}) - \min(A_{i1}, A_{i2}, A_{i3}, A_{i4})$$

$$\text{Range}_{C_i} = \max(C_{i1}, C_{i2}, C_{i3}, C_{i4}) - \min(C_{i1}, C_{i2}, C_{i3}, C_{i4})$$

$$\text{Range}_{E_i} = \max(E_{i1}, E_{i2}, E_{i3}, E_{i4}) - \min(E_{i1}, E_{i2}, E_{i3}, E_{i4})$$

A response should be flagged for adjudication when:

$$\text{Range}_{A_i} \geq 1$$

or:

$$\text{Range}_{C_i} \geq 1$$

or:

$$\text{Range}_{E_i} \geq 1$$

A response should also be flagged if the coders disagree about whether the response is fully unclear or partially codable:

$$\max(U_{i1}, U_{i2}, U_{i3}, U_{i4}) - \min(U_{i1}, U_{i2}, U_{i3}, U_{i4}) = 1$$

or:

$$\max(P_{i1}, P_{i2}, P_{i3}, P_{i4}) - \min(P_{i1}, P_{i2}, P_{i3}, P_{i4}) = 1$$

Flagged cases should be sent to the adjudicator model using the same codebook. The adjudicator should assign final racial Attribution, racial Color-blindness, racial Explicitness, unclear-response flag, and partial coding flag values by applying the codebook directly, not by averaging the prior model responses.

15 Disagreement Flag

A response should be flagged for adjudication when coders disagree substantially on either the racial Attribution or racial Color-blindness dimensions. A response is flagged when:

$$\max(A_{i1}, A_{i2}, A_{i3}) - \min(A_{i1}, A_{i2}, A_{i3}) \geq 1$$

or:

$$\max(C_{i1}, C_{i2}, C_{i3}) - \min(C_{i1}, C_{i2}, C_{i3}) \geq 1$$

A response may also be flagged when the models disagree substantially about racial Explicitness:

$$\max(E_{i1}, E_{i2}, E_{i3}) - \min(E_{i1}, E_{i2}, E_{i3}) \geq 1$$

A response should also be flagged if coders disagree about whether the response is fully unclear or partially codable:

$$\max(U_{i1}, U_{i2}, U_{i3}) - \min(U_{i1}, U_{i2}, U_{i3}) = 1$$

or:

$$\max(P_{i1}, P_{i2}, P_{i3}) - \min(P_{i1}, P_{i2}, P_{i3}) = 1$$

Flagged cases should be sent to an adjudicator model using the same codebook. The adjudicator should assign final racial Attribution, racial Color-blindness, racial Explicitness, unclear-response flag, and partial coding flag values by applying the codebook directly, not by averaging the prior model responses.

16 Reliability and Validation

Because no human-coded gold standard is available, the LLM-coded measure should be evaluated through inter-model reliability, prompt stability, and construct validation.

16.1 Inter-Model Reliability

Reliability should be assessed separately for each rACE dimension and for the response-quality flags. Report the following:

- exact agreement across LLM coders for each rACE dimension;
- within-one-category agreement across LLM coders;
- pairwise weighted Cohen's kappa for racial Attribution, racial Color-blindness, and racial Explicitness;
- Krippendorff's alpha for racial Attribution, racial Color-blindness, and racial Explicitness;
- agreement on the unclear-response flag;
- agreement on the partial coding flag;
- percentage of responses flagged for adjudication.

16.2 Prompt Stability

Recode a random subset of responses using alternative versions of the coding prompt. Report:

- correlation between original-prompt and alternative-prompt scores for racial Attribution, racial Color-blindness, and racial Explicitness;
- mean absolute difference across prompts for racial Attribution, racial Color-blindness, and racial Explicitness;
- agreement on the unclear-response flag across prompt versions;
- agreement on the partial coding flag across prompt versions;
- weighted kappa across prompt versions;
- whether the estimated treatment effect changes across prompt versions.

16.3 Construct Validity

Assess whether the three rACE scores behave as expected in relation to existing survey variables.

Higher racial Attribution scores, indicating more individual or group-blame attribution, should be positively associated with individual-blame explanations for racial inequality, cultural-deficiency explanations, opposition to government assistance for racial minorities, and belief that minorities are responsible for their own disadvantage.

Higher racial Color-blindness scores, indicating more color-blind race interpretation, should be positively associated with racial resentment, symbolic racism, denial of discrimination, opposition to race-conscious policy, belief in meritocracy, conservative ideology, Republican partisanship, and reverse-racism or White-victimhood claims.

Higher racial Explicitness scores, indicating more explicit racial content, should not be interpreted as inherently more or less color-blind. Instead, racial Explicitness should be used to distinguish whether racial reasoning is implicit, mixed, or explicit.

The unclear-response flag and partial coding flag should be analyzed separately. Main analyses can be estimated using fully codable responses and then re-estimated including partially codable responses where the relevant dimension is available. For visual analyses, responses missing racial Explicitness can still be plotted if racial Attribution and racial Color-blindness are available; these responses should be marked with a distinct visual indicator for missing explicitness.

17 Implementation Workflow

The rACE coding procedure should be implemented in stages. Before coding the full set of open-ended responses, I first draw a random 10% pilot sample from the full dataset. The pilot sample is used to evaluate whether the prompt, output format, disagreement rules, adjudication procedure, and aggregation rules function as intended.

17.1 Step 1: Create a Random 10% Pilot Dataset

The full dataset contains 2,000 respondents. I draw a random 10% pilot sample, corresponding to 200 open-ended responses. The pilot sample includes both substantive and non-substantive responses so that the unclear-response flag can be evaluated during the pilot stage.

The open-ended response is stored in `q30_emo`. The respondent identifier is stored in `caseid`. To prevent the LLM coders from using respondent-level information, the file provided to the LLMs contains only two columns: `caseid` and `response_text`. A separate full-data version of the same sampled cases is retained for later merging with respondent covariates, treatment condition, partisanship, race, weights, and other variables.

The 10% pilot sample is generated in R as follows:

```
library(dplyr)
library(stringr)

# Load full dataset
awp <- read.csv("awp.csv")

# Create response checks
df_check <- awp %>%
  mutate(
    response_text = as.character(q30_emo),
```

```
    response_trim = str_squish(response_text),
    has_text = !is.na(response_trim) & response_trim != "",
    obvious_nonsubstantive = case_when(
      !has_text ~ FALSE,
      str_detect(
        str_to_lower(response_trim),
        "^(don'?t know|dk|idk|not sure|no opinion|none|n/a|na|no|nothing|unsure|\\.)$"
      ) ~ TRUE,
      TRUE ~ FALSE
    ),
    likely_substantive = has_text & !obvious_nonsubstantive
  )

# Draw random 10% pilot sample
set.seed(1234)

pilot_sample <- df_check %>%
  slice_sample(prop = 0.10)

# Blinded file for LLM coding: only ID and response text
pilot_blind <- pilot_sample %>%
  select(caseid, response_text = response_trim)

write.csv(
  pilot_blind,
  "rACE_pilot_random_10pct_blind.csv",
  row.names = FALSE
)

# Full file for later merging and analysis: all variables retained
pilot_full_for_later <- awp %>%
  semi_join(
    pilot_sample %>% select(caseid),
    by = "caseid"
  )

write.csv(
  pilot_full_for_later,
  "rACE_pilot_random_10pct_full_for_later.csv",
  row.names = FALSE
)

# Check output dimensions
nrow(pilot_blind)
ncol(pilot_blind)
```

```
nrow(pilot_full_for_later)
ncol(pilot_full_for_later)
```

This procedure produces two files. The first file, `rACE_pilot_random_10pct_blind.csv`, is used for LLM coding and contains only `caseid` and `response_text`. The second file, `rACE_pilot_random_10pct_full_for_later.csv`, is retained for post-coding analysis and contains all original variables for the sampled respondents.

17.2 Step 2: Run Independent LLM Coding

Each pilot response should be independently coded by four LLMs using identical coding instructions. The initial coding models are:

- GPT-4.1;
- Claude Sonnet 4.6;
- Claude Opus 4.6 Reasoning;
- Llama-4-Maverick-17B.

The adjudication model should not be used in the initial coding stage. GPT-5.5 Reasoning should be reserved for adjudication of cases with substantial inter-model disagreement.

Each model should return a CSV file with one row per response and the following columns:

```
caseid
response_text
model_name
race_attribution
race_colorblindness
race_explicitness
unclear_flag
partial_coding_flag
quadrant
confidence
primary_dimension
secondary_dimensions
rationale
flags
```

Each model's output should be saved separately. For example:

```
rACE_pilot_GPT41.csv
rACE_pilot_ClaudeSonnet46.csv
rACE_pilot_ClaudeOpus46Reasoning.csv
rACE_pilot_Llama4Maverick17B.csv
```

The four model output files should then be combined into a long-format dataset, with one row per response-model combination.

17.3 Step 3: Validate Model Outputs

Before aggregating scores, all model outputs should be checked for formatting and coding validity. The validation procedure should confirm that:

- the output is a valid CSV file;
- each input row appears exactly once in each model's output;
- `caseid` and `response_text` are preserved;
- `model_name` is filled in correctly for each model;
- `race_attribution`, `race_colorblindness`, and `race_explicitness` take only the values -1, -0.5, 0, 0.5, 1, or `null`;
- `unclear_flag` and `partial_coding_flag` take only the values 0 or 1;
- cases with `unclear_flag = 1` have `null` values for all three rACE dimensions;
- cases with `partial_coding_flag = 1` have at least one valid rACE score and at least one `null` rACE score;
- cases missing either racial Attribution or racial Color-blindness are assigned `quadrant = not plotted`;
- cases with valid racial Attribution and racial Color-blindness but missing racial Explicitness are retained for plotting and flagged as missing explicitness.

Invalid or malformed outputs should be rerun using the same prompt before aggregation.

17.4 Step 4: Assess Inter-Model Disagreement

For each response, calculate disagreement across the four initial LLM coders separately for each rACE dimension. For each dimension, compute the range across non-missing model scores:

$$\text{Range}_{A_i} = \max(A_{i1}, A_{i2}, A_{i3}, A_{i4}) - \min(A_{i1}, A_{i2}, A_{i3}, A_{i4})$$

$$\text{Range}_{C_i} = \max(C_{i1}, C_{i2}, C_{i3}, C_{i4}) - \min(C_{i1}, C_{i2}, C_{i3}, C_{i4})$$

$$\text{Range}_{E_i} = \max(E_{i1}, E_{i2}, E_{i3}, E_{i4}) - \min(E_{i1}, E_{i2}, E_{i3}, E_{i4})$$

A response should be flagged for adjudication when:

$$\text{Range}_{A_i} \geq 1$$

or:

$$\text{Range}_{C_i} \geq 1$$

or:

$$\text{Range}_{E_i} \geq 1$$

A response should also be flagged if the LLM coders disagree about whether the response is fully unclear or partially codable:

$$\max(U_{i1}, U_{i2}, U_{i3}, U_{i4}) - \min(U_{i1}, U_{i2}, U_{i3}, U_{i4}) = 1$$

or:

$$\max(P_{i1}, P_{i2}, P_{i3}, P_{i4}) - \min(P_{i1}, P_{i2}, P_{i3}, P_{i4}) = 1$$

17.5 Step 5: Adjudicate Flagged Cases

Only flagged cases should be sent to the adjudication model. Adjudication should be performed by GPT-5.5 Reasoning, which is not used in the initial coding stage.

The adjudicator should receive:

- the original open-ended response;
- the rACE codebook;
- the four initial model outputs;
- the four model rationales;
- the adjudication instructions.

The adjudicator should assign final scores by applying the codebook directly. It should not average or defer to the prior model outputs. The adjudicator should return final values for racial Attribution, racial Color-blindness, racial Explicitness, the unclear-response flag, the partial coding flag, quadrant, confidence, rationale, and flags.

17.6 Step 6: Aggregate Final Scores

For responses not flagged for adjudication, final rACE scores should be assigned using the median score across the four initial LLM coders. The median should be computed separately for each dimension over non-missing values:

$$A_i^{final} = \text{median}(A_{i1}, A_{i2}, A_{i3}, A_{i4})$$

$$C_i^{final} = \text{median}(C_{i1}, C_{i2}, C_{i3}, C_{i4})$$

$$E_i^{final} = \text{median}(E_{i1}, E_{i2}, E_{i3}, E_{i4})$$

If fewer than two coders provide a valid score for a dimension, the final value for that dimension should remain missing unless adjudicated.

For the unclear-response flag, the final value should be assigned using majority rule for unflagged cases:

$$U_i^{final} = \begin{cases} 1 & \text{if at least three of four coders assign } U = 1 \\ 0 & \text{otherwise} \end{cases}$$

For the partial coding flag, the final value should also be assigned using majority rule for unflagged cases:

$$P_i^{final} = \begin{cases} 1 & \text{if at least three of four coders assign } P = 1 \\ 0 & \text{otherwise} \end{cases}$$

For flagged cases, the adjudicator's final scores and flags should replace the median model scores and majority-rule flags.

17.7 Step 7: Conduct Researcher Review of the Pilot

After the pilot coding is complete, the researcher should manually inspect a small set of coded responses. This review is not used as a human-coded gold standard. Instead, it is used to identify systematic prompt or codebook problems before coding the full dataset.

The review should focus on whether:

- models distinguish racial Explicitness from racial Color-blindness;
- models avoid coding general conservatism as color-blindness unless it denies, minimizes, naturalizes, or rationalizes racial inequality;
- reverse-racism and White-victimhood claims are consistently identified;
- structural explanations are distinguished from individual or group-blame explanations;
- mixed responses are distinguished from fully unclear responses;
- partially codable responses are correctly flagged;
- responses with missing racial Explicitness but valid racial Attribution and racial Color-blindness remain eligible for visualization.

If systematic problems are identified, the prompt and codebook should be revised before coding the full dataset. Any revisions should be documented.

17.8 Step 8: Produce Pilot Visualizations

After the pilot scores are aggregated, produce a proof-of-concept rACE visualization. The main visualization should plot:

$$x_i = C_i^{final}$$

$$y_i = A_i^{final}$$

with racial Explicitness represented by grayscale intensity when E_i^{final} is available. Responses with valid A_i^{final} and C_i^{final} but missing E_i^{final} should be plotted using a hollow or outlined point. Responses with missing A_i^{final} or C_i^{final} should not be plotted on the main Cartesian plane but may be summarized separately.

Additional respondent characteristics, such as partisanship or treatment condition, may be represented using point shape or separate facets. These respondent characteristics are merged only after LLM coding is complete and are not shown to the LLM coders.

17.9 Step 9: Scale to the Full Dataset

The full dataset should be coded only after the pilot confirms that the prompt, output format, disagreement rules, adjudication process, and visualization procedure are functioning as intended.

For the full coding run, the researcher should save:

- raw model outputs;
- cleaned and validated model outputs;
- adjudication inputs;
- adjudication outputs;
- final aggregated rACE scores;
- response-quality flags;
- model names and coding dates;
- prompt versions;
- code used to parse, validate, aggregate, and visualize the data.

This produces a reproducible audit trail for the LLM coding procedure.

18 Methods Paragraph

Open-ended responses were coded using the rACE coding scheme, a theory-derived three-dimensional approach for classifying open-ended racial discourse. The name rACE highlights that "racial" is the common modifier across the three dimensions: racial Attribution, racial Color-blindness, and racial Explicitness. The coding scheme draws on scholarship conceptualizing color-blind racism as an ideology that denies, minimizes, naturalizes, or rationalizes racial inequality through frames such as abstract liberalism, naturalization, cultural racism, and minimization of racism. It also incorporates the distinction between color-evasion and power-evasion in the psychological literature on color-blind racial ideology, as well as the distinction between explicit and implicit racial communication in research on racial appeals.

Rather than coding responses on a single linear scale, each response was assigned three scores. Each dimension was coded using a five-point ordinal scale taking values -1, -0.5, 0, 0.5, and 1. The racial Attribution score ranges from structural explanation to individual or group-blame explanation. The racial Color-blindness score ranges from color-conscious to color-blind reasoning. The racial Explicitness score ranges from implicit or coded racial meaning to explicit racial meaning. The racial Attribution and racial Color-blindness scores locate each response on a Cartesian plane, while the racial Explicitness score captures whether the racial meaning is implicit, mixed, or explicit. When racial Explicitness could not be coded but racial Attribution and racial Color-blindness could be coded, the response remained eligible for the Cartesian visualization and was marked using a distinct visual indicator for missing explicitness. The coding procedure also includes an unclear-response flag identifying blank, irrelevant, nonsensical, or otherwise impossible-to-classify responses, and a partial coding flag identifying responses for which at least one rACE dimension can be coded but one or more others cannot.

Before coding the full dataset, I drew a random 10% pilot sample from the full set of open-ended responses using R. The sample was drawn using `slice_sample(prop = 0.10)` after setting a random seed `set.seed(1234)`. The blinded pilot file provided to the LLMs contained only the respondent identifier, `caseid`, and the open-ended response text. Respondent covariates, treatment assignment, partisanship, race, weights, and other respondent characteristics were withheld from the LLM coders. A separate full-data version of the same sampled cases was retained for later merging and analysis.

Given the need for rapid coding and the absence of available human coders, I use LLM-based coding rather than traditional manual annotation. This choice is also supported by recent work showing that large language models can perform well on social-scientific text annotation tasks. For example, Törnberg (2025) finds that large language models outperform both expert coders and supervised classifiers in annotating political social media messages. However, I do not treat LLM coding as automatically valid. The coding procedure therefore uses multiple independent LLM coders, structured outputs, disagreement flags, adjudication, prompt-stability checks, inter-model reliability statistics, and construct validation against related survey measures.

Each response was independently coded by four large language models using identical coding instructions: *GPT-4.1*, *Claude Sonnet 4.6*, *Claude Opus 4.6 Reasoning*, and *Llama-4-Maverick-17B*. Final scores were assigned using the median model score for each rACE dimension. Cases of substantial inter-model disagreement were flagged for adjudication.

Adjudication was performed by a fifth model, *GPT-5.5 Reasoning*, which was not used in the initial coding stage and was instructed to apply the codebook directly rather than average or defer to the prior model outputs. Reliability was assessed using inter-model agreement, weighted kappa, Krippendorff's alpha, prompt-stability checks, and agreement on the response-quality flags. Construct validity was assessed by examining associations between the LLM-coded dimensions and related survey measures.

References

Bonilla-Silva, Eduardo. 2017. *Racism without Racists: Color-Blind Racism and the Persistence of Racial Inequality in America*. Old Saybrook, Conn.: Tantor Media.

Mendelberg, Tali. 2001. *The race card: campaign strategy, implicit messages, and the norm of equality*. Princeton, N.J.: Princeton University Press.

Neville, Helen A, Germine H Awad, James E Brooks, Michelle P Flores and Jamie Bluemel. 2013. "Color-Blind Racial Ideology: Theory, Training, and Measurement Implications in Psychology." *The American psychologist* 68(6):455–466.

Törnberg, Petter. 2025. "Large language models outperform expert coders and supervised classifiers at annotating political social media messages." *Social Science Computer Review* 43(6):1181–1195.