

# Large Language Models Naively Recover Ethnicity from Individual Records

Noah Dasanaïke\*

*First Version: January 13, 2026*

*This Version: January 30, 2026*

*[Click here for the latest version](#)*

## Abstract

I demonstrate that large language models can infer ethnicity from names with accuracy exceeding that of Bayesian Improved Surname Geocoding (BISG) without additional training data, enabling inference outside the United States and to contextually appropriate classification categories. Using stratified samples from Florida and North Carolina voter files with self-reported race, LLM-based classification achieves up to 84.7% accuracy, outperforming BISG (68.2%) on balanced samples. I test six models including Gemini 3 Flash, GPT-4o, and open-source alternatives such as DeepSeek v3.2 and GLM-4.7. Enabling extended reasoning can improve accuracy by 1–3 percentage points, though effects vary across contexts; including metadata such as party registration reaches 86.7%. LLM classification also reduces the income bias inherent in BISG, where minorities in wealthier neighborhoods are systematically misclassified as White. I further validate using Lebanese voter registration with religious sect (64.3% accuracy), Indian MPs from reserved constituencies (99.2%), and Indian land records with caste classification (74.0%). Aggregate validation across India, Uganda, Nepal, Armenia, Chile, and Costa Rica using original full-count voter rolls demonstrates that the method recovers known population distributions where naming conventions are distinctive. For large-scale applications, small transformer models fine-tuned on LLM labels exceed BISG accuracy while enabling local deployment at no cost.

---

\*PhD Candidate, Department of Government, Harvard University

# 1 Introduction

Individual-level data on race, ethnicity, and other ascriptive identities are essential for studying questions across the social sciences, such as political representation and voting behavior, yet such information is frequently unavailable in administrative records. I use “ethnicity” broadly hereafter to encompass race, caste, religious sect, and related ascriptive categories. Researchers studying these and related questions often possess only names and geographic information, requiring them to infer demographic characteristics for downstream analysis. The prevailing approach for such inference in the United States is Bayesian Improved Surname Geocoding (BISG), introduced by Imai and Khanna (2016), which combines Census surname-race distributions with geographic demographic data to generate probabilistic race predictions.

BISG has been largely rejected in research on the United States because it produces low accuracy and uninterpretable probability estimates. The method faces fundamental limitations, however, in broader application. Bayesian models such as BISG require training data for classification, in this case relying on Census or voter file data linking surnames to racial distributions. Such data exists almost exclusively in this context, where privacy regulations on individual data are less restrictive. BISG is also constrained to the racial categories defined by the US Census, making it unsuitable for studying religion, caste, or ethnic categories that differ from official administrative categories. And BISG typically uses only surnames, ignoring potentially informative first and middle names. The consequences of these constraints include limited applicability of name-based inference outside the United States, reduced accuracy for groups where full names carry demographic signal, and dependence on official categorization schemes that may not match research questions.

I propose an alternative approach using large language models (LLMs). These models, trained on sizable text corpora, implicitly learn associations between names and demographic characteristics across diverse cultural contexts. By prompting an LLM to classify names into specified ethnic categories, either alone or with additional meta information (such as arbitrarily defined geographic location) one can obtain predictions without any labeled training data. This enables ethnicity inference in countries where no Census surname data exists and for categorization schemes that researchers define to match their research questions. Many countries either lack the administrative infrastructure to frequently conduct censuses, legally prohibit the collection of racial and ethnic data entirely, as in France, or do not do so for political sensitivity reasons, as in Turkey. The LLM approach circumvents these barriers by drawing on associations learned from publicly available text rather than requiring government-

collected training data at granular geographic levels.

I validate the approach through individual-level comparison with BISG in the United States using Florida and North Carolina voter files that include self-reported race, through validation in Lebanon where voter registration records include religious sect, and through validation in India using Members of Parliament from reserved constituencies and land records with caste labels. I also conduct aggregate validation by comparing inferred ethnic distributions to census data across six additional countries, five of which I collect original full-count voter rolls in order to do so: India, Uganda, Nepal, Armenia, Chile, and Costa Rica. I test six models: Gemini 3 Flash, the cost-efficient version of the current state-of-the-art proprietary model; GPT-4o, representing a now-dated but widely-used proprietary alternative; GPT-4.1-mini, a smaller and cheaper OpenAI model; and, following recent arguments that open-source models are an ethical imperative for social science (Spirling, 2023), DeepSeek v3.2 (685B), GLM-4.7 (355B) and GLM-4.7-Flash (30B), the former two being state-of-the-art open-source models (with publicly available weights) and the last a smaller variant. The results demonstrate that LLM-based classification achieves accuracy comparable to or exceeding BISG in the United States while offering substantially greater flexibility for comparative research, though performance varies considerably across geographic contexts and classification categories.

## 2 Related Work

Prior work on name-based ethnicity inference comprises two main approaches: Bayesian methods that combine surname distributions with geographic priors, and machine learning approaches that learn character-level patterns from labeled corpora. BISG (Imai and Khanna, 2016), belonging to the former category, remains the dominant approach in the United States. The method combines Census surname-race distributions with block-level demographic data using Bayes' rule to generate probabilistic race predictions. Subsequent work has addressed limitations in the original formulation. Imai et al. (2022) introduce fully Bayesian BISG (fBISG), which accounts for measurement error in Census surname counts and incorporates first and middle name information, improving accuracy particularly for racial minorities. Greengard and Gelman (2025) demonstrate that statistical dependence of surname and geolocation within racial categories produces biases for minority subpopulations, introducing a raking-based calibration that augments BISG with surname-by-geolocation distributions from voter files. Argyle and Barber (2024) show that BISG misclassification rates correlate with demographic and socioeconomic factors, with

racial minorities in wealthy, educated, and politically active neighborhoods most likely to be misclassified as white. Machine learning approaches, conversely, train classifiers on labeled name-ethnicity datasets. Ethnicolor (Chintalapati et al., 2018), for instances, use neural networks trained on Florida voter file data, achieving approximately 75% macro-precision on US racial categories. Rethnicity (Xie, 2022) employs bidirectional LSTMs with class rebalancing to improve minority group accuracy. NamePrism (Ye et al., 2017) applies naive Bayes to 74 million name-nationality pairs from email and social media data, enabling classification across 39 nationality groups. These methods achieve comparable accuracy to BISG within the United States but share its fundamental limitation: they require labeled training data for each target population and classification scheme. Large language models offer a potential solution to these limitations. Books, news and research articles, and the various other textual corpora that constitute the underlying training data for these models frequently mention individuals by name alongside ethnic, religious, or national identifiers.

### 3 Method

The approach is simple. For each record, I construct a prompt requesting classification into a constrained set of categories appropriate to the local context. A typical prompt takes the form: “Classify the race/ethnicity of this person based on their education and income. education: [education]. income: [income]. Return only one of: [categories].” Geography varies by context, defined at the county level for the American voter file inputs and at the region or province level elsewhere. For output classification categories, I use state-specific groups for the American voter files, religious sects for Lebanon, and ethnic or caste categories as defined in the respective census for comparison in other countries. I set temperature to zero for deterministic outputs.

I use the six proprietary and open-source models earlier listed to assess performance variation across model families and sizes. I implement BISG using the `wru` R package (Khanna et al., 2024) and run it first with surname plus county-level demographics from the 2020 Census, then with surname only. For all methods, I assign individuals to the category with the highest predicted probability. To provide a conservative benchmark of LLM capabilities, I use simple prompts without optimization, employing no prompt engineering, few-shot examples, or iterative refinement. Performance improvements are likely achievable through prompt optimization, particularly in comparative contexts where providing examples of typical names for each category could improve classification of ambiguous cases. In the Appendix, I show that accuracy ranges from

75.7% to 81.3% across six prompt variants on a Florida subsample.

For validation using the Florida and North Carolina voter files, I take stratified samples of 10,000 records from each data set (2,500 per racial category). Both states record self-reported race in voter registration, providing ground truth labels. One might be concerned that voter files could appear in LLM training data; this is plausible for North Carolina, where the voter file is publicly available online, but less likely for Florida, where the file must be formally requested and is not freely distributed. I also use a stratified sample of 1,000 records from the replication data of Lee and Velez (2025), which assembles a dataset of local elected officials with verified race labels to benchmark image-based methods for predicting race and ethnicity. For the Lebanon validation, I use a stratified sample of 3,500 records from Lebanese voter rolls (500 per religious sect), which includes both Arabic names (thus confirming that performance does not rely on Latin script) and religious sect; the Lebanese voter file is not available online.

For aggregate validation, I compare inferred ethnic distributions to aggregate census data across six additional countries: India (2019), Uganda (2025), Nepal (2025), Armenia (2017), Chile (2024), and Costa Rica (2025). For each country, I use a sample of 500 records from original full count electoral rolls, classify names into locally appropriate ethnic or religious categories, and compare aggregate proportions to the most recent census composition data. With the exception of India, the voter rolls for the five other countries come from scraping public electoral commission websites or are recovered from online archives, and represent to my knowledge the first cross-national use of modern full-count voter files. These aggregate comparisons cannot assess individual-level accuracy but provide evidence on whether the method implicitly recovers known population-level distributions.

## 4 Results

Table 1 presents overall accuracy for each method, model, and input configuration in Florida and North Carolina using stratified samples of 10,000 records (2,500 per racial category). With full names and geography, Gemini achieves 83.8% accuracy in Florida and 83.5% in North Carolina, compared to 68.2% and 68.9% for BISG with surname and county. The BISG figures are lower than typically reported because stratified sampling equalizes racial categories rather than reflecting population proportions; BISG performs well on White voters but poorly on minorities. Ablating input components reveals the relative contribution of first names and geography: removing geography (full name only) reduces GPT-4o accuracy by 0.6pp in Florida, while removing first names (sur-

name + geo) reduces accuracy by 8.7pp, indicating that first names carry substantial additional signal beyond surnames.

Method	Florida						North Carolina					
	Gem.	4o	4.1-m	DS	4.7	4.7F	Gem.	4o	4.1-m	DS	4.7	4.7F
Full name + geo	<b>83.8</b>	78.5	70.5	76.2	74.5	62.3	<b>83.5</b>	80.5	74.1	78.4	77.9	72.3
Full name only	<b>82.8</b>	78.2	69.7	71.2	75.0	70.7	<b>82.3</b>	79.6	72.9	73.6	77.2	73.7
Surname + geo	<b>76.0</b>	69.9	65.9	67.7	61.8	58.0	<b>74.3</b>	69.0	66.8	68.1	63.6	59.7
Surname only	<b>69.7</b>	64.7	58.4	63.7	60.4	59.0	<b>68.3</b>	64.5	60.6	64.4	61.6	60.7
BISG (surname + geo)	68.2						68.9					

Table 1: Overall accuracy (%) by input configuration. Gem. = Gemini 3 Flash; 4o = GPT-4o; 4.1-m = GPT-4.1-mini; DS = DeepSeek v3.2; 4.7 = GLM-4.7; 4.7F = GLM-4.7-Flash. BISG uses surname + county geography via the wru package. n=10,000 per state (2,500 per racial category).

Table 2 presents recall by racial category for Florida and North Carolina, revealing for which racial categories the methods differ most. LLMs show substantial improvements for Black voters, with Gemini achieving 81.7% recall in Florida and 80.0% in North Carolina compared to 48.8% and 53.1% for BISG, likely owing to the latent contextual signal provided by first and middle names that is unavailable to surname-only methods. Hispanic voters are classified with over 90% recall by LLM methods in Florida and over 84% in North Carolina, reflecting distinctive surname patterns for that group.

Race	Florida							North Carolina						
	Gem.	4o	4.1-m	DS	4.7	4.7F	BISG	Gem.	4o	4.1-m	DS	4.7	4.7F	BISG
White	88.4	88.5	90.6	88.7	<b>90.9</b>	70.1	86.8	84.4	88.2	<b>90.5</b>	89.6	86.4	83.5	85.0
Black	<b>81.7</b>	72.2	48.9	61.4	59.3	43.1	48.8	<b>80.0</b>	68.3	52.1	59.3	66.4	60.6	53.1
Hispanic	92.3	91.6	91.4	<b>92.6</b>	90.3	94.9	88.8	84.9	<b>85.7</b>	83.8	85.6	81.6	78.4	77.0
Asian	<b>72.8</b>	61.6	51.2	62.4	57.3	40.5	48.5	<b>84.5</b>	79.6	70.2	79.4	77.4	67.4	60.5
Average	<b>83.8</b>	78.5	70.5	76.3	74.5	62.3	68.2	<b>83.5</b>	80.5	74.2	78.5	78.0	72.3	68.9

Table 2: Recall by racial category (%), full name + geography condition. Gem. = Gemini 3 Flash; 4o = GPT-4o; 4.1-m = GPT-4.1-mini; DS = DeepSeek v3.2; 4.7 = GLM-4.7; 4.7F = GLM-4.7-Flash. n=10,000 per state (2,500 per racial category).

I also apply LLM classification to the replication data from Lee and Velez (2025), which compares multiple methods (including image-based) for predicting the race and ethnicity of local elected officials. Using a random sample of 1,000 records from their test set, Gemini with geographic context achieves 90.4% accuracy, exceeding both their BISG benchmark (84.5%) and their hybrid approach combining images and names (88.8%). GPT-4o achieves 88.3%, comparable to the hybrid approach, while GLM-4.7 achieves 82.8%, slightly below BISG. Their hybrid approach requires collecting facial photographs, potentially raising privacy concerns and limiting generalized applicability, while the LLM approach

requires only names and supports any supplementary meta data (such as geography). Note that accuracy on the Lee and Velez data appears higher than on the stratified Florida and North Carolina samples in Table 1; this reflects the demographic composition of their dataset (approximately 65% White), as White names are generally easier to classify correctly.

Method	Gem.	4o	DS	GLM-4.7	4.7F
LLM (full name + city)	<b>90.4</b>	88.3	82.6	82.8	60.0
LLM (full name only)	81.4	<b>83.0</b>	72.9	76.4	80.5
<i>Lee &amp; Velez benchmarks:</i>					
Hybrid (image+name)			88.8		
fBISG			85.4		
BISG			84.5		
Surname-only			81.9		

Table 3: Comparison with Lee and Velez (2025) benchmarks (%). Gem. = Gemini 3 Flash; 4o = GPT-4o; DS = DeepSeek v3.2; 4.7F = GLM-4.7-Flash. n=1,000 random sample from test set.

The Lebanese voter file affords a rare opportunity for individual-level validation outside the United States. Religious sect is both recorded in voter registration and associated with naming patterns. Overall accuracy for the Lebanon sample is 64.3% for Gemini 3 Flash, 54.5% for GPT-4o, 44.8% for GPT-4.1-mini, 44.5% for DeepSeek v3.2, 47.7% for GLM-4.7, and 31.2% for GLM-4.7-Flash (n=3,500). The LLMs achieve high accuracy for groups with distinctive naming conventions: Armenian Orthodox (whose names are highly distinctive, up to 97.4% for Gemini), Shia (whose names often include religious markers like Hussein and Ali), and Maronites (whose names frequently have French-influenced forms). Performance drops for Roman Orthodox and Roman Catholic, possibly reflecting overlap in naming conventions with other Christian sects. GLM-4.7-Flash struggles substantially with this task, achieving reasonable accuracy only for Shia and Maronite classifications.

India’s reserved constituency system provides another opportunity for individual-level validation outside the United States. Under Indian electoral law, certain parliamentary constituencies are reserved for candidates from Scheduled Castes (SC, historically disadvantaged Dalit communities) or Scheduled Tribes (ST, indigenous Adivasi communities). Members of Parliament elected from these constituencies must by law belong to the designated category, providing legally verified ground truth for caste/tribe classification. Using the 130 MPs from reserved constituencies in the 18th Lok Sabha (2024), I test whether LLMs can distinguish SC from ST members based on names alone. Gemini 3 Flash

Sect	Gem.	4o	4.1-mini	DS	4.7	4.7F
Armenian Orthodox	<b>97.4</b>	95.6	73.4	64.0	81.4	15.4
Shia	86.4	<b>90.4</b>	81.0	76.8	78.0	70.4
Maronite	<b>77.0</b>	71.6	34.6	62.0	66.8	66.2
Sunni	76.8	54.0	<b>82.2</b>	81.6	74.6	55.4
Druze	<b>55.2</b>	45.8	26.8	5.6	6.0	0.2
Roman Orthodox	<b>45.2</b>	14.0	5.8	10.6	19.2	0.0
Roman Catholic	<b>12.2</b>	10.2	10.0	10.8	8.0	10.6
Overall	<b>64.3</b>	54.5	44.8	44.5	47.7	31.2

Table 4: Accuracy by religious sect (%), Lebanon. Gem. = Gemini 3 Flash; 4o = GPT-4o; 4.1-mini = GPT-4.1-mini; DS = DeepSeek v3.2; 4.7 = GLM-4.7; 4.7F = GLM-4.7-Flash. n=3,500 (500 per sect).

achieves 99.2% accuracy (129/130), with 98.8% recall for SC and 100% for ST. GPT-4o achieves 91.5%, GLM-4.7 achieves 86.9%, and GLM-4.7-Flash achieves 73.8%. The smaller open-source model struggles specifically with ST classification (29.8% recall), potentially reflecting limited representation of tribal naming patterns in its training data.

Category	Gem.	4o	DS	GLM-4.7	4.7F
Scheduled Caste (SC)	<b>98.8</b>	94.0	50.6	88.0	98.8
Scheduled Tribe (ST)	<b>100.0</b>	87.2	91.5	85.1	29.8
Overall	<b>99.2</b>	91.5	65.4	86.9	73.8

Table 5: Accuracy for SC/ST classification (%), India Lok Sabha MPs from reserved constituencies. Gem. = Gemini 3 Flash; 4o = GPT-4o; DS = DeepSeek v3.2; 4.7F = GLM-4.7-Flash. n=130 (83 SC, 47 ST).

To test classification across all four caste reservation categories—SC, ST, OBC (“other backward class”), and general (“unreserved”)—I draw on two additional datasets with ground truth labels: elected village council heads (*sarpanch*) from the state of Rajasthan, which have official reservation category labels, and Bihar land records where caste (*jati*) is recorded alongside owner names.<sup>1</sup> Performance varies substantially between datasets (Table 6). For a stratified random sample with Gemini 3 Flash, Bihar achieves 74.0% overall accuracy compared to 57.0% for Sarpanch. Why the difference? The accuracy of these inferences depends heavily on whether names encode caste (or any other ascriptive identity), which in India varies by region and community. Bihar names frequently

<sup>1</sup>Data provided by Gaurav Sood.

include explicit caste surnames: Paswan, Chamar, and Dusadh, for instance, are well-known SC surnames, and overall accuracy for this subgroup is 97.3%. We also see relatively (80.0%) strong classification of unreserved groups. In contrast, Sarpanch names from Rajasthan lack consistent caste markers. Interestingly, classification of scheduled tribes actually shows the opposite pattern: Sarpanch achieves 72.0%, while Bihar falls to 58.7%.

Category	Gemini 3 Flash		DeepSeek v3.2		GLM-4.7	
	Sarpanch	Bihar	Sarpanch	Bihar	Sarpanch	Bihar
Scheduled Caste (SC)	<b>48.0</b>	<b>97.3</b>	42.7	74.7	4.0	56.0
Scheduled Tribe (ST)	72.0	<b>58.7</b>	57.3	30.7	<b>73.3</b>	52.0
Other Backward Class (OBC)	57.3	60.0	<b>60.0</b>	<b>78.7</b>	18.7	25.3
General (unreserved)	50.7	80.0	40.0	37.3	<b>85.3</b>	<b>92.0</b>
Overall	<b>57.0</b>	<b>74.0</b>	50.0	55.3	45.3	56.3

Table 6: Accuracy for four-way caste classification (%). Sarpanch: elected village heads with official category labels (Rajasthan). Bihar: land records with recorded *jati*. Bold indicates best performance per dataset. n=300 each (75 per category, stratified).

I also conduct aggregate validation by comparing inferred ethnic distributions to census data across six additional countries (Table 7). These comparisons draw on original full count voter files obtained through scraping public electoral commission websites or accessing online archives. For each country, I take a random sample of the full voter rolls, then classify individuals using the respective census category and then aggregate the proportions classified into each group. Performance varies substantially depending on the distinctiveness of naming conventions. The strongest results appear in India, where Gemini’s inferred Hindu and Muslim shares fall within 0.5 percentage points of census figures, and in Armenia, where distinctive naming patterns yield near-perfect estimates. Both Gemini and GPT-4o discriminate well among minority groups: in Uganda, estimates fall within 3.3 percentage points for all 10 ethnic groups, and in Nepal, the best model achieves the closest estimate for 10 of 11 categories. However, performance appears to degrade when naming conventions overlap or minority populations have limited representation in the training data. Despite the highlighted performance among minority ethnic groups, Uganda nevertheless shows systematic over-prediction of the largest three ethnic groups. Nepal presents the most significant challenges, though the proprietary models substantially reduce bias: for instance, GPT-4o predicts 19.2% Chhetri and Gemini 20.8% (vs 16.6% census), a marked improvement over GLM-4.7’s 31.7%. Gemini also accurately identifies Yezidis in Armenia (1.0% vs 1.2% census), a small minority that GPT-4o and the open-source GLM models miss entirely.

Indigenous populations in Costa Rica remain difficult to detect, with Gemini identifying only 0.2% versus the 2.4% census figure, as naming patterns likely overlap with the majority (that is, they are likely not mutually-exclusive with other groups).

Country	Category	Census	Gem.	4o	DS	GLM-4.7	4.7F	Err.
India	Hindu	79.8	<b>80.2</b>	81.0	78.6	79.0	82.0	0.4
India	Muslim	14.2	<b>13.9</b>	13.1	12.0	13.2	9.5	0.3
India	Christian	2.3	2.1	0.8	1.0	<b>2.4</b>	2.1	0.1
India	Sikh	1.7	<b>2.2</b>	3.0	2.8	3.6	2.4	0.5
India	Buddhist	0.7	<b>0.2</b>	0.2	0.0	0.2	0.1	0.5
India	Jain	0.4	0.3	<b>0.4</b>	0.9	0.2	0.3	0.1
India	Other	0.9	<b>1.0</b>	1.4	4.7	1.3	3.5	0.1
India	<i>Avg. error</i>	–	<b>0.3</b>	0.9	1.5	0.7	1.6	0.3
Uganda	Baganda	16.5	24.7	<b>19.8</b>	29.3	28.0	51.5	3.3
Uganda	Banyankole	9.6	17.1	11.6	<b>11.2</b>	15.4	13.8	1.6
Uganda	Basoga	8.8	10.5	<b>7.8</b>	11.6	9.9	7.6	1.0
Uganda	Iteso	7.0	<b>8.1</b>	4.0	9.1	11.3	0.0	1.1
Uganda	Langi	6.1	<b>7.1</b>	8.6	4.6	4.9	0.1	1.0
Uganda	Bagisu	4.9	5.4	2.0	<b>4.6</b>	4.2	9.7	0.3
Uganda	Acholi	4.4	4.8	<b>4.4</b>	7.3	6.3	2.5	0.0
Uganda	Lugbara	4.2	3.4	3.2	5.4	<b>4.9</b>	1.8	0.7
Uganda	Batooro	3.2	<b>3.6</b>	2.8	5.3	5.5	10.5	0.4
Uganda	Banyoro	3.0	<b>3.1</b>	2.0	5.3	4.3	1.1	0.1
Uganda	<i>Avg. error</i>	–	2.2	<b>1.7</b>	3.0	3.1	7.2	1.7
Nepal	Chhetri	16.6	20.8	<b>19.2</b>	22.2	31.7	64.0	2.6
Nepal	Brahmin	12.2	16.2	15.8	<b>11.7</b>	13.8	17.6	0.5
Nepal	Magar	7.1	6.1	<b>6.4</b>	5.9	12.5	3.5	0.7
Nepal	Tharu	6.6	<b>7.6</b>	8.2	7.6	2.2	0.9	1.0
Nepal	Tamang	5.8	<b>6.1</b>	5.4	4.7	4.0	0.1	0.3
Nepal	Newar	5.0	<b>5.7</b>	6.2	6.4	2.4	0.0	0.7
Nepal	Kami	4.8	6.2	<b>4.8</b>	5.2	3.8	0.8	0.0
Nepal	Yadav	4.0	<b>3.6</b>	5.6	15.3	10.0	7.5	0.4
Nepal	Rai	2.3	<b>3.2</b>	3.4	4.7	4.1	3.4	0.9
Nepal	Gurung	2.0	<b>2.3</b>	2.8	6.3	2.4	0.2	0.3
Nepal	Limbu	1.5	<b>1.5</b>	1.2	1.3	1.2	0.9	0.0
Nepal	<i>Avg. error</i>	–	<b>1.3</b>	<b>1.3</b>	2.7	3.7	7.6	1.3
Armenia	Armenian	98.1	<b>98.2</b>	99.2	97.5	99.1	100.0	0.1
Armenia	Yezidi	1.2	<b>1.0</b>	0.0	0.0	0.2	0.0	0.2
Armenia	Russian	0.5	0.8	0.8	0.0	<b>0.4</b>	0.0	0.1
Armenia	<i>Avg. error</i>	–	<b>0.2</b>	0.9	0.8	0.7	1.2	0.2
Chile	Mestizo	87.2	90.0	82.2	<b>89.5</b>	82.7	96.0	2.3
Chile	Mapuche	9.9	4.0	5.2	6.3	<b>9.6</b>	0.4	0.3
Chile	European	1.6	<b>2.6</b>	5.0	3.9	6.1	1.0	1.0
Chile	Aymara	0.7	1.8	2.0	0.0	<b>1.2</b>	0.1	0.5
Chile	<i>Avg. error</i>	–	2.7	3.6	<b>2.2</b>	2.5	4.9	2.2
Costa Rica	Mestizo	83.6	92.6	72.0	75.6	<b>83.0</b>	98.7	0.6
Costa Rica	White	8.7	3.2	<b>9.2</b>	16.5	9.8	0.4	0.5
Costa Rica	Afro-Carib.	7.8	3.6	8.4	<b>7.3</b>	6.0	0.0	0.5
Costa Rica	Indigenous	2.4	<b>0.2</b>	0.0	0.2	0.0	0.1	2.2
Costa Rica	Asian	0.2	0.4	0.4	0.3	0.5	<b>0.2</b>	0.0
Costa Rica	<i>Avg. error</i>	–	4.2	3.1	3.7	<b>1.2</b>	6.7	1.2

Table 7: Aggregate validation (%) against census data. Bold indicates closest to census. Gem. = Gemini 3 Flash; 4o = GPT-4o; DS = DeepSeek v3.2; 4.7F = GLM-4.7-Flash; Err. = minimum error. n=500–1,000 per country.

A key advantage of LLM-based classification over BISG is the ability to incorporate additional information. BISG uses only surname and geographic demographics; extending it to other covariates would require specifying new conditional probability distributions and obtaining appropriate training data. LLMs face no such constraint: any information expressible in natural language can be included in the prompt. To assess the added value of additional metadata, I test six input configurations on a Florida subsample (n=300) using Gemini 3 Flash: (1) name and county (baseline), (2) name, county, and age, (3) name, county, and party registration, (4) name and zip code (replacing county with finer geography), (5) name, county, and gender, and (6) all available features combined. Age may inform classification because first name conventions vary across generations. Party registration correlates with race in the United States, though including it raises questions about endogeneity if the goal is to study racial differences in political behavior. Zip codes provide finer geographic resolution than county but may not be consistently represented in the training data. Gender might plausibly resolve ambiguity for names used across ethnic groups with different gender associations.

Input Configuration	Accuracy	vs. Baseline
BISG (surname + county)	79.8%	-4.9pp
Name + County (baseline)	84.7%	—
Name + County + Age	85.0%	+0.3pp
Name + County + Party	<b>86.7%</b>	+2.0pp
Name + Zip Code	84.7%	0.0pp
Name + County + Gender	84.0%	-0.7pp
All features	84.7%	0.0pp

Table 8: Accuracy with additional metadata (%), Florida subsample. Baseline uses full name and county; all features includes name, county, city, zip code, age, gender, and party registration. n=300.

Table 8 presents results. Party registration provides the largest improvement (+2.0pp), with Gemini here achieving 86.7% accuracy, though again, researchers should be cautious when classification features are correlated with downstream outcomes. Age yields a modest gain (+0.3pp). Zip code provides no improvement over county-level geography, and gender slightly reduces accuracy (-0.7pp). Combining all features fails to improve over the baseline. These results demonstrate that LLMs can leverage metadata that BISG cannot incorporate, though gains depend on the informativeness of features in context.

## 5 Scaling

For large-scale applications, such as population-level inference, researchers can use LLM classification or manual labeling to create a training set, then fine-tune a lightweight model on these labels for cheaper local inference. Table 9 presents results from knowledge distillation experiments using Gemini-labeled training data, fine-tuning small transformer models with LoRA (Low-Rank Adaptation) across three datasets. Each dataset is split 70/20 for training and evaluation: Florida (n=,200 train, 300 test), North Carolina (n=1,200 train, 00 test), and Lebanon (n=8000 train, 200 test). I use a small observation count to reflect practical application for researchers (i.e., hand labeling). Teacher accuracy reflects Gemini’s performance on the held-out test set.

Model	Dataset	Teacher	Base	Fine-tuned	Gap
Qwen3-4B	Florida	83.3%	72.0%	78.7%	-4.7pp
Qwen3-1.7B	Florida	83.3%	60.3%	<b>79.7%</b>	-3.7pp
Qwen3-0.6B	Florida	83.3%	7.3%	75.7%	-7.7pp
Qwen3-4B	North Carolina	71.7%	63.7%	<b>68.0%</b>	-3.7pp
Qwen3-1.7B	North Carolina	71.7%	64.0%	67.7%	-4.0pp
Qwen3-0.6B	North Carolina	71.7%	18.0%	<b>68.0%</b>	-3.7pp
Qwen3-4B	Lebanon	75.5%	15.5%	<b>65.0%</b>	-10.5pp
Qwen3-1.7B	Lebanon	75.5%	28.0%	58.0%	-17.5pp
Qwen3-0.6B	Lebanon	75.5%	14.5%	52.0%	-23.5pp

Table 9: Knowledge distillation results. Fine-tuned transformers use LoRA on Gemini-labeled data. Base shows zero-shot performance before fine-tuning. Gap is relative to teacher (Gemini 3 Flash). 80/20 train/test split.

Fine-tuning small transformer models on Gemini labels substantially improves over base model performance and approaches teacher accuracy on US data. This is particularly impressive given the small training label count. On Florida, Qwen3-1.7B achieves 79.7% accuracy after fine-tuning, exceeding BISG’s 76.2% while enabling local deployment. Performance transfer varies by context: North Carolina shows similar retention, while Lebanon exhibits larger gaps likely reflecting weaker Arabic name representations in the base models. These distilled models can run on consumer hardware, eliminating API or server costs associated with proprietary or large-parameter open source alternatives.

## 6 Reasoning

Recent large language models offer configurable “thinking” or “reasoning” modes that allow the model to perform intermediate reasoning steps before generating a final answer. These modes trade increased computation time and token usage for potentially improved accuracy on complex tasks. I evaluate whether enabling extended reasoning improves ethnicity classification accuracy using Gemini 3 Flash, which offers thinking levels from “minimal” to “high,” and two Qwen models (235B and 8B parameters), which provide an explicit reasoning toggle. I also test Gemini 3 Pro, the larger and more expensive variant of the Gemini 3 family, which requires reasoning to be enabled and does not support minimal thinking mode. This last model has been excluded from benchmarking thus far owing to cost and time inefficiency for practical use by researchers.

Model	Reasoning	Florida		North Carolina	
		Accuracy	Gain	Accuracy	Gain
Gemini 3 Flash	Minimal	83.7%	—	<b>82.3%</b>	—
Gemini 3 Flash	High	<b>84.7%</b>	+1.0pp	82.0%	−0.3pp
Gemini 3 Pro	High	83.3%	—	79.3%	—
Qwen 3 235B	Off	72.3%	—	77.0%	—
Qwen 3 235B	On	74.3%	+2.0pp	75.7%	−1.3pp
Qwen 3 8B	Off	64.3%	—	68.0%	—
Qwen 3 8B	On	67.7%	+3.4pp	70.0%	+2.0pp

Table 10: Effect of reasoning on classification accuracy. n=300 per state (full name + geography). Gemini uses thinking levels; Qwen uses explicit reasoning toggle. Gemini 3 Pro requires reasoning enabled.

Table 10 presents results for Florida and North Carolina samples. In Florida, enabling extended reasoning improves accuracy across all models, with gains inversely related to base model capability: Gemini 3 Flash gains 1.0 percentage point, Qwen 3 235B 2.0 percentage points, and Qwen 3 8B 3.4 percentage points. In North Carolina, the results are more mixed: Gemini 3 Flash drops 0.3 percentage points and Qwen 3 235B drops 1.3 percentage points, while Qwen 3 8B still gains 2.0 percentage points. This inconsistent result across states and models suggests that extended reasoning may not reliably improve classification, and that gains observed in one context are unlikely to generalize. Reasoning also substantially increases inference time (approximately  $2.8\times$  for Gemini and  $5\times$  for Qwen) and token consumption. Notably, Gemini 3 Pro with high reasoning (83.3%) performs slightly worse than Gemini 3 Flash with high reasoning (84.7%), despite being substantially more expensive. For most applications, the

minimal or no reasoning setting is likely to prove sufficient; however, it can improve performance when running smaller models locally.

## **7 Income Bias**

Argyle and Barber (2024) document that BISG misclassification rates correlate with socioeconomic factors: racial minorities in wealthier neighborhoods are systematically more likely to be misclassified as White. As noted earlier in this paper, inference bias such as this raises concerns for downstream analyses, where classification errors may be endogenous to the outcome quantities of interest.

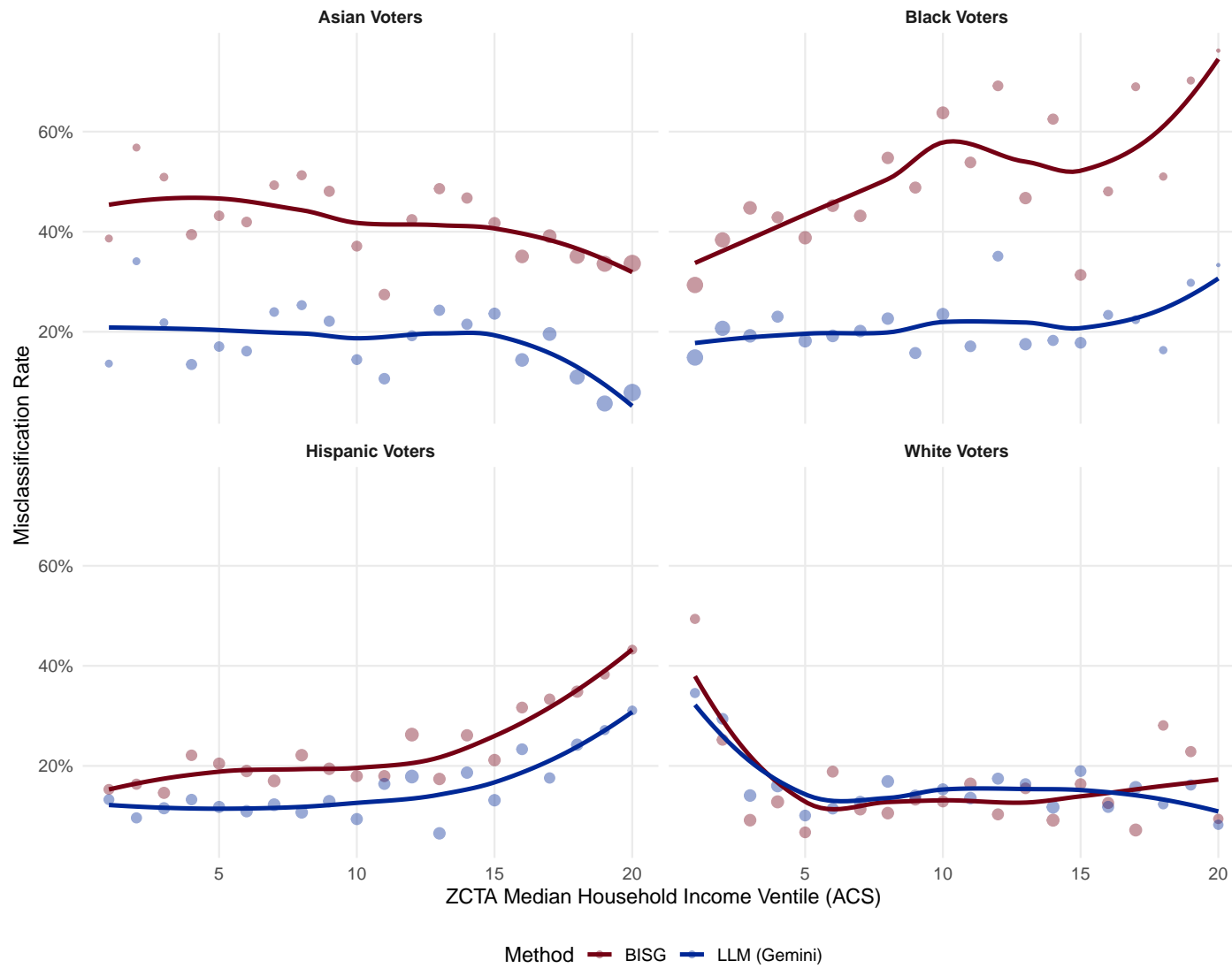


Figure 1: Misclassification rate by ZCTA median household income (ACS 5-year estimates) and race, North Carolina voters. Points show average misclassification rate for each income ventile (5% quantile), sized by number of observations. Lines show loess fits. For Black voters, BISG misclassification increases sharply with neighborhood income; LLM shows no such gradient.

I replicate their analysis using North Carolina voter records matched to Census Bureau median household income by ZIP Code Tabulation Area (ZCTA) from the American Community Survey 5-year estimates (Figure 1). Following their methodology, I calculate median household incomes at a low-level geographic level (here ZCTA), then plot misclassification rates against the geographic median separately faceted by race. For Black voters, BISG misclassification rises from approximately 40% in the lowest income ventile to over 65% in the highest, a pattern consistent with Argyle and Barber’s finding that minorities in wealthier areas are systematically misclassified as White. Regressing misclassification on ZCTA median income for Black voters, the BISG coefficient is 0.037 ( $p < 0.001$ ), indicating a 3.7 percentage point increase in misclassification per \$10,000 of neighborhood income. A similar regression for Gemini is not statistically significant ( $p = 0.06$ ). This differential bias likely arises because BISG relies on geographic racial composition as a prior, thus conflating residence in a predominantly White area with being White, whereas LLMs incorporate geographic metadata in a less strictly defined manner. Interestingly, both models become less biased with local income for Asian voters, and more biased with Hispanic voters.

## 8 Discussion

The LLM approach addresses several limitations of BISG that have thus far constrained empirical research on ethnic politics outside the United States. Because these models require no labeled training data, researchers can apply them to any country or categorization scheme without obtaining prior surname-geographic distributions. The approach also allows researchers to leverage information otherwise discarded by BISG or related methods: first and middle names that carry substantial ethnic signal, geographic information at multiple hierarchical levels, and additional metadata such as sex, age, or, party registration, the last of which improved accuracy by 2 percentage points in Florida (Table 8). Neither language nor script of origin appeared to impede model performance in the validation exercises, thus further removing constraints otherwise imposed by the need for pre-processing.

Nonetheless, several limitations warrant consideration. Updates to proprietary models may affect reproducibility, though open-source models with fixed weights or fine-tuned small models address this concern. LLMs may encode societal biases about name-ethnicity associations learned from training data, which may disadvantage individuals with atypical names for their group and affect downstream analyses. Privacy concerns are reduced where open-source models are

deployed locally. Researchers may also be concerned about sensitivity to prompt formulation; testing six prompt variants on a Florida subsample (n=300), accuracy ranges from 75.7% (minimal) to 81.3% (detailed), a spread of 5.6 percentage points, suggesting the method is reasonably robust (see Appendix for details).

As with any method, researchers should validate before deployment by hand-coding a small sample or comparing inferred distributions to available benchmarks. Warning signs include implausible concentrations in a single category and poor minority group performance. Researchers should also verify that classification errors are not systematically correlated with variables of interest in downstream analyses. The results also reveal that model selection involves a trade-off between accuracy and reproducibility: Gemini 3 Flash achieves the highest accuracy across most validation exercises, but proprietary models accessed via API may change over time and require sending data to external servers. GLM-4.7 achieves comparable though slightly lower accuracy while offering fixed weights, local deployment, and long-term reproducibility. For applications where results must be verifiable and replicable, open-source or researcher fine-tuned models are likely preferable despite the accuracy penalty.

In summary, large language models provide a viable alternative to BISG for ethnicity inference from names, achieving comparable or superior accuracy without requiring additional labeled data. The method extends naturally to classification schemes that BISG cannot address, including religious sect, caste, and ethnicity, and to countries lacking detailed surname-geographic probability distributions. Performance varies by context depending on how strongly naming conventions correlate with group membership and the extent to which these conventions are mutually exclusive across groups. Overall, the method opens new possibilities for studying ethnic politics in contexts where traditional methods cannot be applied.

## References

Argyle, Lisa P., and Michael Barber. "Misclassification and Bias in Predictions of Individual Ethnicity from Administrative Records." *American Political Science Review* 118, no. 2 (2024): 1058–1066.

Chintalapati, Rajashekar, Suriyan Laohaprapanon, and Gaurav Sood. "Predicting Race and Ethnicity from the Sequence of Characters in a Name." arXiv preprint arXiv:1805.02109 (2018).

Greengard, Philip, and Andrew Gelman. "A Calibrated BISG for Inferring Race from Surname and Geolocation." *Journal of the Royal Statistical Society Series A: Statistics in Society* (2025).

Imai, Kosuke, and Kabir Khanna. "Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records." *Political Analysis* 24, no. 2 (2016): 263–272.

Imai, Kosuke, Santiago Olivella, and Evan T. R. Rosenman. "Addressing Census Data Problems in Race Imputation via Fully Bayesian Improved Surname Geocoding and Name Supplements." *Science Advances* 8, no. 49 (2022): eadc9824.

Khanna, Kabir, Brandon Bertelsen, Santiago Olivella, Evan Rosenman, Alexander Rossell Hayes, and Kosuke Imai. "wru: Who are You? Bayesian Prediction of Racial Category Using Surname, First Name, Middle Name, and Geolocation." R package version 3.0.3 (2024).

Lee, Diana Da In, and Yamil R. Velez. "Measuring Descriptive Representation at Scale: Methods for Predicting the Race and Ethnicity of Public Officials." *British Journal of Political Science* 55, e110 (2025).

Spirling, Arthur. "Why Open-Source Generative AI Models Are an Ethical Way Forward for Science." *Nature* 616, no. 7957 (2023): 413.

Xie, Fangzhou. "Rethnicity: An R Package for Predicting Ethnicity from Names." *SoftwareX* 17 (2022): 100965.

Ye, Junting, Shuchu Han, Yifan Hu, Baris Coskun, Meizhu Liu, Hong Qin, and Steven Skiena. "Nationality Classification Using Name Embeddings." In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1897–1906 (2017).