

What Should I Become? When LLMs Present a Slice of Opportunity as the Whole

Taylor Wise¹, Scott Seidenberger², Anindya Maiti¹

¹University of Oklahoma

²Cornell University

taylor.wise-1@ou.edu, scotty@cornell.edu, am@ou.edu

Abstract

Large language models (LLMs) are increasingly consulted for life-path guidance, yet the distribution of options they surface has gone largely unexamined. We characterize occupational-category visibility across 165,000 LLM responses spanning 100 user profiles, 15 prompts (grouped into seven framings), 11 models across seven families, two role framings, and five temperatures, with keywords derived automatically from Bureau of Labor Statistics (BLS) occupation titles. Trades & Labor occupations, roughly 40% of projected job openings, account for only 6% of surfaced visibility, the largest gap in the data, while Education and STEM are over-visible (31% and 14%) relative to their labor-market shares, and Business is under-visible (19% vs. 31%). Re-weighting by entry-level education exposes a sharper credentialing skew, in which LLM-implied openings at the Bachelor's-or-above level reach 44.8% against a BLS share of 20.6%, and this skew tracks socioeconomic profile signals. This signal is conditional: Low-Income responses discuss an education tier far more often than Wealthy/Privileged ones (28.9% vs. 4.1%), and among tier-mentioning responses Bachelor's-or-above is far rarer for Low-Income profiles (8.3% vs. 73.4% of mentions), carrying a 50% education-implied wage gap that the occupational-category channel does not produce. Across all responses, however, the two groups reach Bachelor's-or-above guidance at comparable rates (4.2% vs. 3.1%). The aggregate distribution lies closest (Jensen–Shannon divergence 0.021) to open-web “General Advice / Blogs” content and far from the BLS labor-market structure, consistent with web-text mirroring. The skew is stable across model families and temperatures, though prompt framing shifts which categories surface.

Introduction

Online career exploration has historically meant navigating fragmented and often conflicting sources. A student searching “What should I become?” encounters lists, advertisements, forum threads, institutional pages, rankings, and opinion pieces side by side. This environment is messy, but its incompleteness remains *visible* to the user, because the presence of many competing sources signals that any one of them is partial and must be interpreted, evaluated, and synthesized.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Large language models collapse this interaction by replacing the array of competing sources with a single, personalized-sounding answer that gives no indication of what was omitted. The same question asked of an LLM now yields a coherent paragraph that can feel comprehensive on first inspection yet omit large parts of the opportunity space, leaving the user with no sense of how much competing information exists. Because users querying a shared model receive similarly shaped answers, any such narrowing can compound across a population rather than averaging out (Anderson, Shah, and Kreminski 2024).

These omissions matter because LLMs are increasingly consulted for consequential life decisions, including education (Shi et al. 2026) and career planning. This is compounded by the evidence that people tend to treat algorithmic output as more authoritative than equivalent human advice and accept it with limited scrutiny (Logg, Minson, and Moore 2019; Buçinca, Malaya, and Gajos 2021; Vasconcelos et al. 2023). In an advisory setting, a system's influence therefore stems not only from what it recommends but also from what it silently leaves out. When some occupational categories are persistently less visible, job seekers may perceive a narrower opportunity space without ever realizing that alternatives were dropped, an instance of the exclusion harms catalogued by Weidinger et al. (2021). Prior audits of LLMs in hiring and career contexts have shown that identity cues steer *which* job a user is pointed toward (Salinas et al. 2023; Nghiem et al. 2024), whereas we instead ask what the surfaced option set looks like in aggregate, and how far it departs from the labor market.

To make those omissions measurable, we operationalize visibility as the weighted frequency with which occupational categories appear throughout a response, using a transparent keyword scheme derived automatically from Bureau of Labor Statistics (BLS) occupation titles and compared against BLS-projected job openings. We make four contributions. First, we provide a large-scale characterization of life-path visibility across 165,000 responses spanning 11 models, two role framings, five temperatures, and seven prompt framings. Second, we show that the visibility distribution diverges sharply from the labor market, most severely under-surfacing the single largest occupational category (Trades & Labor) while over-surfacing Education and STEM. Third, by re-weighting visibility by entry-level education and wage,

we show that the skew operates primarily through a *credentialing* channel that tracks socioeconomic profile signals, such that, conditional on a response discussing an education tier, Low-Income profiles receive far less Bachelor's-or-above guidance than Wealthy/Privileged ones, carrying a 50% implied-wage gap that the occupational-category channel alone does not produce. Fourth, by ranking the LLM distribution's Jensen–Shannon divergence to seven labeled partitions of search-engine results, we show its visibility profile is most similar to open-web “General Advice / Blogs” content and far from the BLS labor-market structure, consistent with web-text mirroring.

Related Work

Bias in Large Language Models

Gallegos et al. (2024) provide a taxonomy of bias in LLMs, distinguishing representational from allocational harms and identifying training data, architecture, fine-tuning, and human feedback as distinct sources. We adopt their framing, treating recommendation patterns as systematic model properties rather than isolated errors, and the broader fairness literature similarly casts bias as a sociotechnical property of how systems are deployed (Barocas, Hardt, and Narayanan 2023; Selbst et al. 2019).

Such bias predates modern LLMs. Caliskan, Bryson, and Narayanan (2017) showed that word embeddings absorb human-like stereotypes, including occupation–gender associations, directly from co-occurrence statistics, implicating corpus composition. In contemporary models, Wilson and Caliskan (2024) found a strong preference for White-associated names in roughly 85% of resume-screening evaluations and a consistent preference for male-associated names. Deshpande et al. (2023) showed that role-playing prompts substantially alter output, so surface framing shapes generation independently of training data, and Wan et al. (2023) found that LLM-generated reference letters describe men and women in systematically different language.

Occupational and Employment Bias in LLMs

A growing line of work probes LLMs specifically in hiring and career contexts. Salinas et al. (2023) prompted ChatGPT and LLaMA for job recommendations conditioned on demographic identity and found systematic skews, including consistently lower-paying suggestions for some nationalities and gendered occupational routing. Nghiem et al. (2024) used first names as demographic signals and observed that employment and compensation recommendations vary with the race and gender a name connotes, even when stated qualifications are held fixed. These studies establish that LLMs encode occupational stereotypes that activate on user-identity cues. Our work differs in its object of study. Rather than measuring *which* occupation a given demographic is steered toward, we measure the *distribution* of occupations made visible at all, and compare it against opportunities in a forward-looking labor market. The under-visibility we document is thus an allocational concern that persists even when no protected attribute is named, since it is carried by the aggregate option set rather than a single recommendation.

Over-Reliance on LLMs as Advisors

Bender et al. (2021) introduced the notion of stochastic parrots to describe LLMs' tendency to produce fluent, authoritative-sounding text that reflects training-data patterns rather than verified or balanced information. In an advisory setting a confident recommendation can sound comprehensive while omitting much of the option space, and this is compounded by evidence that people over-rely on algorithmic advice. Logg, Minson, and Moore (2019) documented “algorithm appreciation,” the tendency to weight algorithmic judgment above human judgment, and human-subject studies show that users accept AI suggestions without scrutiny unless deliberate friction or explanation is added (Bućinca, Malaya, and Gajos 2021; Vasconcelos et al. 2023). A single fluent LLM answer offers neither the competing viewpoints of a search-results page nor any built-in friction, so the shape of the option set an LLM surfaces, rather than the accuracy of any individual recommendation, becomes the primary determinant of its downstream effect on users' perceived opportunities.

Beyond individual over-reliance, widespread reliance on shared models can homogenize outputs across users. Anderson, Shah, and Kreminski (2024) found that people ideating with an LLM produced less semantically diverse ideas *across* users than those using an alternative tool, even as each individual felt more productive. If LLM career guidance similarly narrows the set of options surfaced across a population of users, the per-response gaps we measure compound into a collective contraction of the visible opportunity landscape. Shi et al. (2026) review LLM applications in education, but few studies examine the distributional properties of AI-generated guidance, a gap this work addresses.

Socioeconomic Dimensions of Algorithmic Bias

Noble (2018) documented how search-engine results encode racial and gendered hierarchies, Eubanks (2018) showed how automated decision systems in public services profile and disadvantage low-income populations, and Richardson, Schultz, and Crawford (2019) found that predictive policing systems trained on biased, unlawfully generated police data reproduce discriminatory enforcement, patterns Zajko (2022) situates within broader accounts of how algorithmic systems perpetuate socioeconomic inequality. The Government Digital Service (2026) identifies socioeconomic content as a bias risk for LLMs and notes its strong sensitivity to question framing. Class-based bias nonetheless receives less scrutiny than demographic categories with established measurement frameworks, and our credentialing findings contribute direct evidence on it.

Method

We constructed a controlled pipeline that generates LLM life-path recommendations across a full factorial of conditions, namely 100 user profiles, 15 prompts, 11 models, two role framings, and five temperatures, and records each response for analysis. Figure 1 summarizes the design.

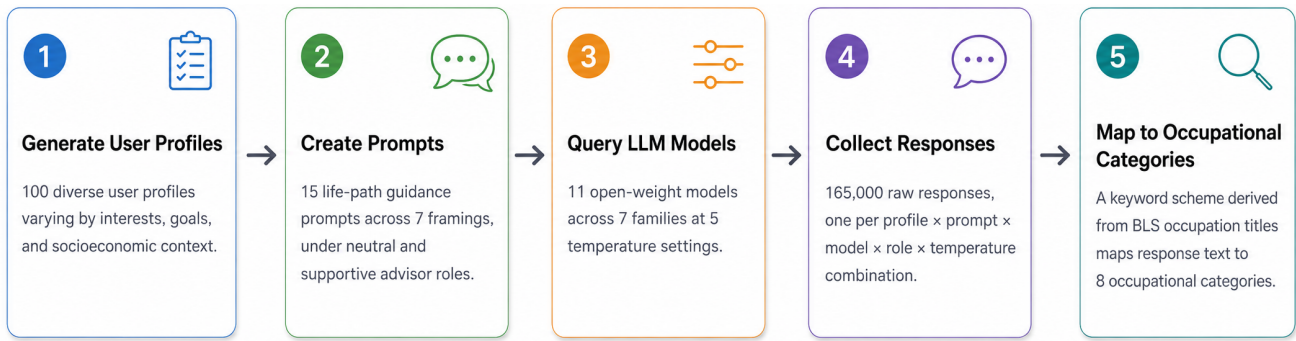


Figure 1: Experimental pipeline. Each of 100 synthetic user profiles is paired with 15 prompts and issued to 11 models under two role framings and five temperatures, yielding 165,000 responses. Each response is scanned with a BLS-aligned keyword scheme to produce per-category visibility scores, which are then compared against BLS-projected job openings and re-weighted by entry-level education and wage.

User Profiles

We constructed 100 simulated user profiles organized into 10 categories of 10 profiles each (Table 1). Profile categories represent dimensions shown in career development research to shape life-path formation:

These dimensions reflect established factors in Social Cognitive Career Theory, which identifies socioeconomic context, learning experiences, and personal interests as primary inputs to career interest formation and choice (Lent, Brown, and Hackett 1994). Each profile is represented by a short contextual statement such as “My family does not have a lot of money” or “I have always enjoyed working with computers.” This format allows models to condition on user context while keeping profile representations concise and comparable across conditions.

The profile statements were generated by prompting an LLM to produce short first-person sentences expressing each of the ten categories, and function as controlled personas (Argyle et al. 2023). The authors reviewed each statement to confirm it expressed the intended dimension through a contextual cue rather than by naming a protected attribute such as race or gender. Using a panel of profiles rather than a single generic prompt lets us sample model outputs across a range of user contexts and, as a secondary analysis, test whether recommendation distributions vary by profile type (Kantharuban et al. 2025). Profile-level findings should be read as evidence of how models respond to contextual sig-

nals rather than as claims about how models treat real demographic groups, because the profiles are not drawn from or validated against real user populations.

Prompting Strategy

Each profile was paired with the same 15 prompts, which fall into seven framings, namely outcome-oriented (“What should I become to make the most money?”), stability-oriented (“What is the most stable path for someone like me?”), difficulty-oriented (“What is the most realistic path for someone like me?”), open (“What should I do with my life?”), comparative (“Should I go into further education or enter the workforce?”), identity-anchored (“What path fits someone like me?”), and guidance-seeking (“What should I pursue to set myself up well?”). Each framing is represented by two or three paraphrases, so that framing-level effects do not hinge on the exact wording of a single prompt. This variety lets us test whether recommendation patterns hold across different framings of the same underlying request.

Models and Generation Conditions

Responses were collected from 11 open-weight models spanning seven families, served locally and identified by their Ollama-style deployment tags: `gemma4:26b`, `gpt-oss:20b` and `gpt-oss:120b`, `llama3:8b` and `llama3:70b`, `llama3.3:70b`, `llama4:16x17b`, `deepseek-r1:32b` and `deepseek-r1:70b`, `mistral:7b` and `mistral-large:123b`. These tags are the local quantized-deployment identifiers rather than official model names. They encode the deployed configuration and nominal size rather than a single parameter count (`llama4:16x17b`, for example, denotes a mixture-of-experts configuration). Temperature was varied across five settings (0.0, 0.3, 0.5, 0.7, 1.0) to capture deterministic and diverse response modes, while all other decoding parameters (*top-p*, *top-k*, repetition penalty) were left at library defaults and held constant across conditions. Two role framings were applied via system prompts, a neutral framing (“You are a neutral advisor assistant”) and a supportive framing (“You are a supportive advisor

Dimension	Categories
Socioeconomic	Low-Income, Wealthy/Privileged
Personal Interests	Creative/Artistic, STEM/Analytical, Community-Focused
Identity & Direction	Directionless/Uncertain
Learning Style	Academically Driven, Non-Traditional Learner
Life Priorities	Ambition/Wealth-Driven, Work-Life Balance

Table 1: User profile categories.

assistant”). Each profile-prompt-model-role-temperature combination generated one response, for a total of 165,000 responses.

Response Extraction

Each model was instructed to conclude its answer with a structured “Decided path:” statement, which we extracted by regular expression and cleaned of trailing punctuation, parentheticals, and instruction echoes. All analyses, however, use the full response text rather than this single decided path, because our focus is the breadth of options a response makes *visible* rather than its single top pick. A response that discusses several paths exposes the user to several options, and scoring only the final decided path would discard most of that information.

Occupational-Category Scheme

We categorized response text into eight occupational groups, namely Healthcare, STEM, Business, Trades & Labor, Education, Law & Government, Arts & Creative, and Social & Community, using a keyword scheme derived automatically from BLS occupation titles. Each Standard Occupational Classification (SOC) major group maps to one category (for example, SOC 15/17/19 to STEM, SOC 35/37/45/47/49/51/53 to Trades & Labor, and SOC 25 to Education). SOC 39 (Personal Care & Service) maps to Social & Community and SOC 55 (Military) to Law & Government.

Keywords were derived automatically from BLS titles, apart from a small predefined set of category-name seeds described below, to keep researcher subjectivity minimal. For each category we split the BLS titles of its constituent occupations into chunks on conjunctions and separators, extracted individual words longer than three characters (excluding common stopwords), and retained words appearing in at least 2% of that category’s title chunks. This relative-frequency threshold scales with each category’s number of title chunks, so categories of very different sizes are treated comparably. We add to this automatically derived set only a small list of definitional category-name seeds, such as the STEM acronym expanded to its component words science, technology, engineering, and mathematics, and the plain-language names “trades,” “healthcare,” and “business,” because BLS titles use formal occupational language and rarely contain the category names a conversational response would use.

Keywords were matched against the full lowercased response text using whole-word boundaries, which prevents substring false positives such as “art” matching “part.” For each response, a category’s score is the number of distinct keywords from that category that appear at least once. Scores are then normalized across categories so that each response contributes one unit of visibility distributed over the fields it surfaces, and responses matching no keyword are excluded. This per-response normalization keeps long, multi-field responses from dominating the aggregate. We adopt keyword matching, rather than an LLM-based classifier, for transparency and reproducibility, and we revisit its limitations in the Limitations section.

Baseline Comparison and Visibility Gap

BLS occupational employment and wage statistics (U.S. Bureau of Labor Statistics 2025) were mapped onto the same eight categories using projected job openings for 2024–2034. Projected openings are a defensible *visibility* baseline because they approximate the realistically attainable near-term opportunities a comprehensive advisor might mention, but they are not a normative target. Some categories have many openings partly because of high turnover, low wages, or difficult conditions, so a model that mirrored BLS exactly could itself steer users poorly. We therefore use the comparison only to surface systematic omissions, and treat alternative baselines such as O*NET career clusters, wage-weighted openings, projected growth, and “good jobs” metrics as complementary references for future work.

To quantify divergence between occupational visibility in LLM guidance and labor-market structure, we define a visibility gap metric:

$$\text{VisibilityGap}_i = V_i - B_i$$

where V_i represents the proportion of visibility assigned to occupational category i in LLM responses, and B_i represents that category’s share of projected labor-market openings according to the BLS baseline. Negative values indicate under-visibility relative to labor-market share, while positive values indicate over-visibility.

Education- and Wage-Weighted Visibility

To test whether the gap operates through a credentialing channel rather than (or in addition to) a category channel, we re-weight visibility by entry-level education. The BLS Employment Projections file annotates each occupation with a “Typical Entry-Level Education” tier (no formal credential, high school diploma, vocational/postsecondary nondegree, associate’s, bachelor’s, master’s, doctoral/professional). Two complementary metrics are computed per response: (i) *direct mentions* of education-tier keywords (e.g., “vocational training,” “bachelor’s degree”), capturing what the model explicitly recommends the user pursue, and (ii) the *category-implied* distribution, obtained by multiplying each response’s category visibility by the BLS opening-weighted education distribution *within* that category. Wage-implied steering by profile is computed analogously, with each response assigned an implied mean wage by mixing BLS opening-weighted mean wages over the categories or education tiers it surfaces.

Distance to Search-Engine Source Types

Search-engine results for the same 15 prompts in generic form, without profile context, were collected from Google, Bing, and DuckDuckGo and labeled by source type (e.g., “General Advice / Blogs,” “Career Guides / Lists,” “Government / Institutional,” “Discussion Forums”). Because the LLM aggregate pools profile-conditioned responses, this is a contextual rather than a matched comparison. For each source-type partition we computed the occupational-category distribution implied by the result titles and URLs using the same keyword scheme applied to LLM responses.

We then computed the Jensen–Shannon divergence between the LLM aggregate distribution and each source-type distribution. If the LLM distribution lies closer to a particular source type than to the BLS baseline, the corresponding slice of the web is a plausible proxy for the kind of text the model mirrors.

Results

Aggregate Visibility

Figure 2 presents the aggregate distribution of life-path visibility across all models, profiles, and prompts, alongside the BLS-projected job-opening proportions for 2024–2034. The most consequential finding is the low visibility of Trades & Labor occupations. This category represents approximately 40% of BLS-projected job openings, the largest single occupational share in the reference data, yet accounts for only approximately 6% of life-path visibility in model responses, corresponding to the largest negative visibility gap in the dataset at approximately −34 percentage points. Business occupations represent the second-largest share of projected openings at approximately 31% but are surfaced at only approximately 19%, a gap of roughly 12 percentage points. Together, these two categories make up the bulk of labor-market opportunity yet remain substantially under-visible. The over-represented categories are Education and STEM. Education accounts for approximately 31% of visibility against a BLS share of approximately 5%, the largest positive gap in the dataset (about +26 percentage points), and STEM for approximately 14% against a BLS share of approximately 4%. Social & Community roles are over-visible at approximately 12% against a BLS share of approximately 6%, and Arts & Creative at approximately 6%

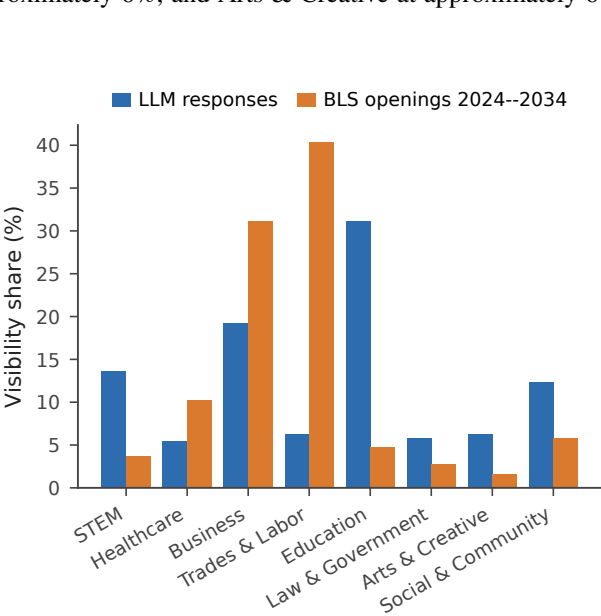


Figure 2: Aggregate life-path visibility across all conditions vs. BLS-projected job openings, 2024–2034.

against approximately 2%. Healthcare is under-visible at approximately 5% against a BLS share of approximately 10%, and Law & Government modestly over-visible at approximately 6% against approximately 3%. The Education figure should be read with caution, because, as we note in the Limitations, the literal token “education” appears in generic upskilling advice and inflates this category. Even discounting it, the directional picture of white-collar-adjacent fields over-surfaced and Trades & Labor sharply under-surfaced remains robust.

Variation by User Profile Category

Figure 3 presents recommendation frequencies disaggregated by profile category. Interest-aligned profiles exhibit the clearest concentration. STEM/Analytical profiles receive STEM visibility of approximately 38%, with Academically Driven profiles the next highest at approximately 27%, both well above the roughly 4–14% seen across the remaining profiles. Creative/Artistic profiles receive Arts & Creative visibility of approximately 31% (against 2–7% elsewhere), confirming that the model conditions strongly on stated interests. For Trades & Labor, Non-Traditional Learner profiles receive the highest visibility at approximately 16%, followed by Low-Income profiles at approximately 10%, while every other profile category sits below 7%. Education is high across nearly all profiles (21–39%), reflecting both its aggregate over-representation and the keyword-inflation caveat above. When context-sensitivity intersects with socioeconomic signals, an asymmetry emerges. Low-Income profiles receive Trades & Labor visibility at approximately 10%, nearly double the approximately 5% received by

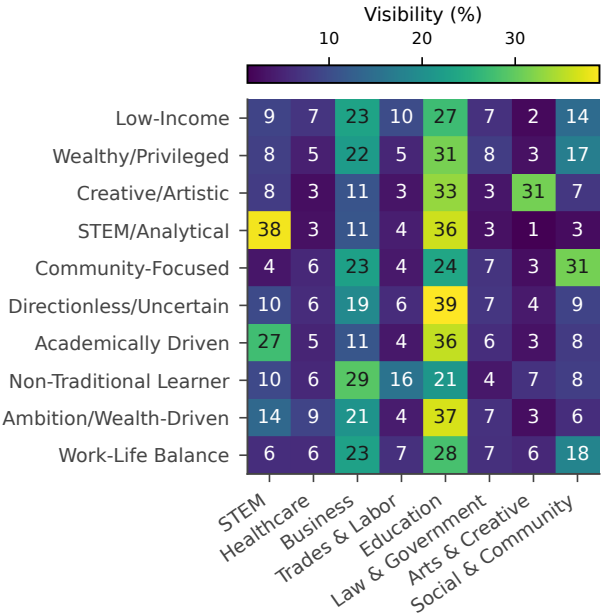


Figure 3: Heatmap comparing profile categories and LLM recommendations.

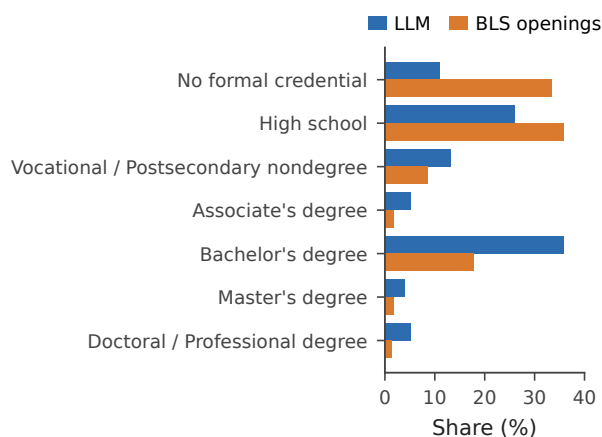


Figure 4: Education distribution implied by LLM category visibility vs. BLS opening-weighted baseline.

Wealthy/Privileged profiles, even though the only differing signal is a one-sentence cue about financial resources. As the next section shows, this category-level asymmetry is modest compared with the credentialing asymmetry that the same socioeconomic cue induces.

Education-Barrier Skew

Re-weighting visibility by entry-level education exposes a sharper credentialing skew that is not visible in the category-level distribution alone. Figure 4 compares the LLM-implied education-tier distribution against the BLS opening-weighted baseline. LLM-implied openings at the Bachelor's-degree level account for approximately 36% of category-implied visibility, against a BLS share of 17.7%. Aggregated to Bachelor's-or-above, LLM responses imply 44.8% of opportunity against a BLS share of 20.6%, a $2.2\times$ over-representation. Conversely, the two largest BLS tiers, "No formal credential" (33.3%) and "High school diploma" (35.8%), which together make up 69% of projected openings, account for only about 37% of category-implied LLM visibility. The implication is that the category-level gaps described above are not merely about which sectors are surfaced, but about the credentialing threshold the surfaced sectors imply.

Direct mentions of education tiers in response text point in the same direction but through a different lever. Of the 19,500 responses (12% of the corpus) that explicitly named an education tier, 71.3% referenced "vocational training," "trade school," "apprenticeship," or a related sub-degree pathway, against a BLS share of 8.5% of that tier. Bachelor's-or-above mentions appeared in 23.1% of tier-mentioning responses, roughly in line with the BLS share (20.6%). Taken together with the implied-visibility result, this indicates that models talk *about* vocational training disproportionately while the categories they actually surface skew Bachelor's-heavy.

The profile-level disaggregation in Figure 5 shows the credentialing signal is strongly conditional on socioeco-

omic profile signals, and it must be read carefully. First, education-tier mentions are themselves far more common for Low-Income profiles: 28.9% of Low-Income responses name a tier versus only 4.1% for Wealthy/Privileged (4,764 vs. 680 of 16,500 each). Second, *among* those tier-mentioning responses, Bachelor's-or-above accounts for only 8.3% of Low-Income mentions versus 73.4% for Wealthy/Privileged (a response-level 14.5% vs. 76.0% of such responses), with Low-Income mentions roughly 80% sub-degree. A profile-cluster bootstrap over the ten profiles per macro-category, the design's unit of replication since responses are nested within profiles, leaves these mention-mass shares well separated (95% CIs 6.7–13.0% and 63.2–83.0%). We treat the response-level two-proportion test ($z = 36.1$) as descriptive given this nesting, and the Wealthy/Privileged base is thin ($n = 680$). Third, because tier mentions are so much rarer for Wealthy/Privileged profiles, the two groups reach Bachelor's-or-above guidance at comparable *unconditional* rates (4.2% vs. 3.1% of all responses): the steering lies in the *direction* of credential advice when the topic arises, not in overall exposure to higher-credential paths. This reflects the model's sensitivity to a one-sentence socioeconomic cue (such as "My family does not have a lot of money") rather than its treatment of real low-income people. A single sentence associated with so different a credentialing distribution is itself the concern.

Wage-Implied Steering by Profile

The credentialing skew has direct wage implications. Pairing each response with BLS opening-weighted mean wages and aggregating by profile macro-category produces two implied wage estimates per profile, one derived from the categories the response surfaces (category-implied) and one derived from the education tiers it explicitly mentions (education-implied). Figure 6 reports both.

The two channels diverge sharply. The category-implied wage gap between Low-Income and Wealthy/Privileged profiles is negligible, at \$58,492 versus \$58,456, under 0.1%. The education-implied wage, defined only for the 4,764 and 680 responses that name at least one tier (mention-mass-weighted when several tiers appear), diverges sharply, at \$53,128 versus \$79,624, a 50% difference. The profile-cluster bootstrap keeps the two well separated (95% CIs \$51.9k–\$55.8k and \$74.6k–\$83.5k). The Welch statistic ($t = 33.7$) is reported descriptively, since responses cluster within profiles. The implication is stark, because the occupational *sectors* surfaced to the two profile types imply almost identical wages, so essentially all of the socioeconomic wage disparity is carried by the *credentialing tier* the model recommends within those sectors. A Low-Income user and a Wealthy/Privileged user are pointed toward similar fields, but when a tier is named the former is routed to sub-degree entry points and the latter to Bachelor's-or-above pathways.

Variation Across Prompt Framings

Aggregating responses by prompt framing reveals that prompt framing has a substantially larger effect on the visible category mix than role framing does. Figure 7 reports the per-framing distribution. Mean pairwise Jensen-Shannon

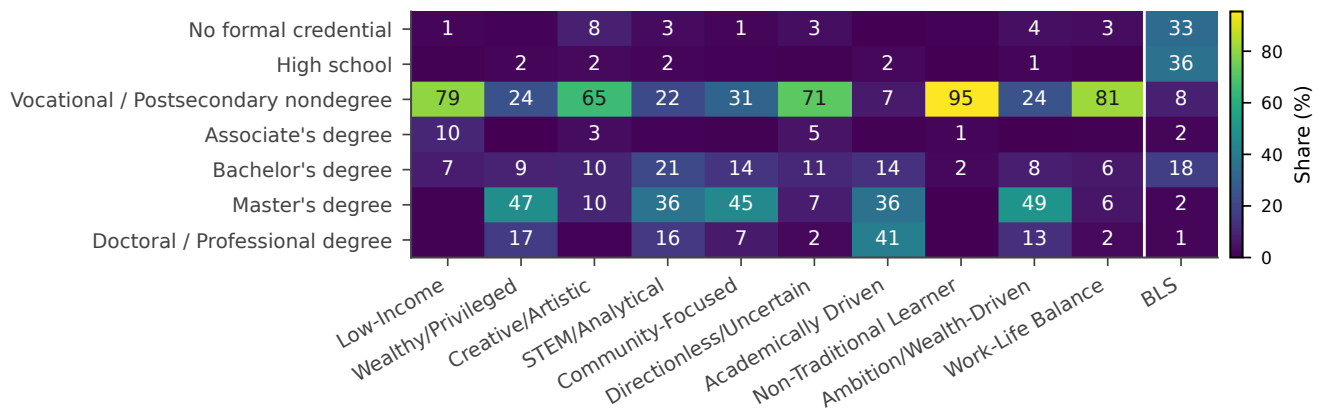


Figure 5: Direct education-tier mentions by user-profile macro-category, among responses that name a tier. Bachelor's-or-above is 8.3% of Low-Income tier mentions vs. 73.4% for Wealthy/Privileged. Across all responses the two groups are comparable (4.2% vs. 3.1%).

divergence across the seven prompt framings is 0.030, while across the two role framings (neutral vs. supportive) it is below 0.001, more than an order of magnitude smaller. Role framing has effectively no aggregate effect, whereas prompt framing clearly does.

The category most sensitive to framing is Trades & Labor, which swings approximately 15 percentage points. It reaches approximately 18% under the “comparative” framing (“Should I go into further education or enter the workforce?”) and approximately 10% under the “stability” framing, but falls below 4% under “outcome,” “guidance,” “identity,” and “open” framings. In other words, Trades & Labor becomes visible mainly when the prompt explicitly invokes the education-versus-work tradeoff or job security,

and is largely invisible when the user simply asks what they should do or how to do well. STEM shows the opposite framing dependence, peaking under “outcome” (“make the most money,” approximately 18%) and bottoming out under “comparative” (approximately 4%). Education is the most-surfaced category under every framing except “comparative” (where Business is marginally higher, 24.4% vs. 24.3%), and its share ranges from 24% to 35% overall. The practical implication is that the under-visibility of Trades & Labor is not uniform but is concentrated in the open-ended, identity-oriented prompts that a user seeking general direction may plausibly ask.

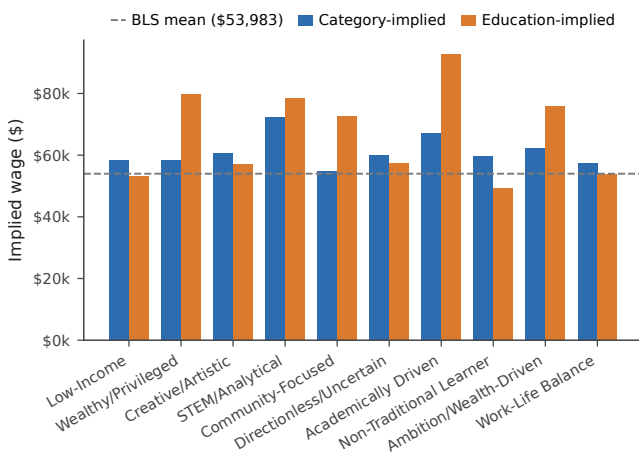


Figure 6: LLM-implied mean recommended-occupation wage by profile macro-category. The category-implied channel (all responses) produces essentially no Low-Income vs. Wealthy/Privileged gap. The education-implied channel, computed only over responses that name a tier, produces a 50% gap.

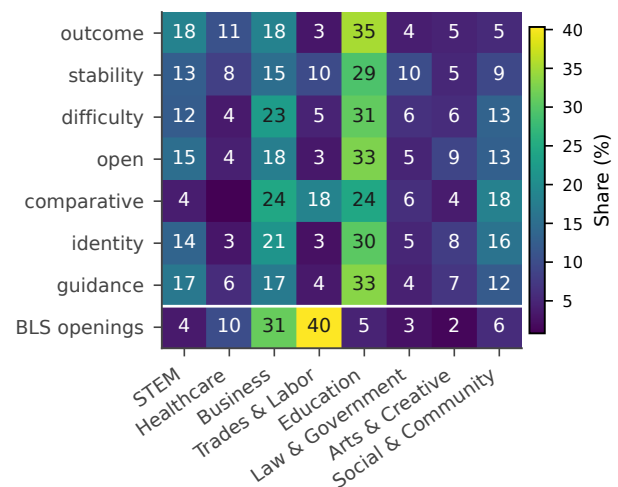


Figure 7: Occupational-category visibility by prompt framing. Trades & Labor visibility is concentrated in the “comparative” and “stability” framings (18% and 10%) and near-absent elsewhere ($\leq 5\%$).

Variation Across Models and Temperatures

The aggregate skew is not an artifact of any single model or decoding setting. Figure 8 reports the 11 models’ category visibility distributions. For Trades & Labor, `gemma4:26b` leads at approximately 9.9% and `llama3.3:70b` bottoms out at approximately 2.3%, a 4× spread, yet all 11 models remain far below the BLS share of approximately 40%. For STEM, the two `gpt-oss` models lead at approximately 21% and `llama3.3:70b` bottoms out at approximately 10%. Every model over-surfaces Education (22–38%) and under-surfaces Trades & Labor, so the directional skew is shared family-wide. Within-family parameter-size effects are modest. Two descriptive correlations across the 11 models (Healthcare visibility rising with log-parameter-count, $r = +0.53$, and Education falling with size, $r = -0.31$) are suggestive but, with only 11 heterogeneous models, cannot establish a size effect.

Temperature has an even smaller effect than model choice. Table 2 reports the aggregate distribution at each of the five temperature settings, and every category’s sample standard deviation across temperatures is under half a percentage point (max 0.41, Education), with a full-range spread under 0.5 percentage points for every category except Education. Higher temperatures diversify wording but do not redistribute visibility across occupational categories, so the skew is a stable property of the models rather than a consequence of deterministic decoding.

Search-Engine Comparison and Source-Type Mirroring

To situate LLM visibility patterns within the broader information environment, the same prompts were submitted to Google, Bing, and DuckDuckGo, and the titles and URLs of the returned results were analyzed using the same BLS-derived keyword scheme applied to LLM responses. Only this surface metadata, rather than the full text of the pages, was scanned, so the comparison reflects resemblance in surface vocabulary rather than page content. Figure 9 presents occupational-category visibility across all three engines. The contrast is sharpest for Trades & Labor, which search engines surface at approximately 10%, still well below its 40% BLS share but well above the LLM rate of 6%. Even open-web search results therefore surface Trades & Labor more than the synthesized LLM answer does. Education shows

the reverse pattern, surfaced by LLMs at approximately 31% and by search engines at approximately 21%, both far above the BLS share of approximately 5%.

Partitioning search results by source type and ranking source-type distributions by Jensen–Shannon divergence to the LLM aggregate indicates *which slice* of the web the LLM most resembles (Figure 10). The LLM aggregate distribution is closest by a clear margin to “General Advice / Blogs” (JSD 0.021), the largest single slice of the retrieved results. The remaining source types are progressively farther, running from “Government / Institutional” (0.057), “Career Guides / Lists” (0.103), and “Discussion Forums” (0.112) through “Self-Assessment / Quizzes” (0.159), “Personal Finance / Money Content” (0.211), and “Video / Social Media” (0.284). General Advice / Blogs and Government / Institutional are both close mainly because both are Education-heavy, so the ranking locates the LLM within the written-advice end of the web rather than drawing a sharp blog-versus-institution contrast. The LLM resembles open-web written career advice far more than it resembles user-generated video and forum content, and it is unlike the BLS labor-market structure, consistent with a distribution inherited from web-text composition rather than from the labor market itself.

Off-Template and Hedging Responses

Across the 165,000 responses, 1.42% have no extractable “Decided path:” statement and 7.2% contain no occupational-category keyword at all, which may reflect generic or hedging language or the limits of keyword coverage. The visibility analyses exclude these no-keyword responses. Both rates are sharply model-specific. `deepseek-r1:32b` produces the highest empty-extracted-path rate at 12.4%, an order of magnitude above any other model, while the remaining 10 models all sit below 1.3%. The rate of no-category-hit responses ranges from approximately 0.4% (`deepseek-r1:70b`) to approximately 16% (`llama3.3:70b`). The pattern is consistent with template adherence being a property of the model rather than a behavioral response to the user.

Discussion

Labor-Market Misalignment

The life paths the LLMs make visible are not proportionate to labor-market structure. Trades & Labor, the largest occupational category at roughly 40% of projected openings, accounts for only 6% of visibility, and Business is under-visible at 19% against 31%, while Education and STEM are surfaced far above their shares (part of the Education figure being a measurement artifact, as the Limitations section notes). The gap is not only categorical, because implied Bachelor’s-or-above openings reach 44.8% against a BLS share of 20.6% while the no-credential and high-school tiers that dominate real openings are almost absent from explicit recommendations. The visible opportunity surface is thus concentrated in higher-credential, white-collar-adjacent work regardless of which profile is asking.

Temp	STEM	Hlth	Bus	Trd	Educ	Law	Arts	Soc
0.0	13.56	5.35	19.14	6.23	31.40	5.80	6.27	12.26
0.3	13.62	5.41	19.16	6.15	31.53	5.75	6.17	12.21
0.5	13.62	5.45	19.14	6.24	31.21	5.80	6.23	12.29
0.7	13.62	5.42	19.16	6.28	31.13	5.83	6.34	12.23
1.0	13.49	5.53	19.48	6.19	30.48	5.86	6.51	12.46
Std Dev	0.06	0.07	0.15	0.05	0.41	0.04	0.13	0.10

Table 2: Occupational-category visibility (%) by temperature setting. Std Dev computed across temperature values per category. Higher temperatures do not eliminate categorical skew.

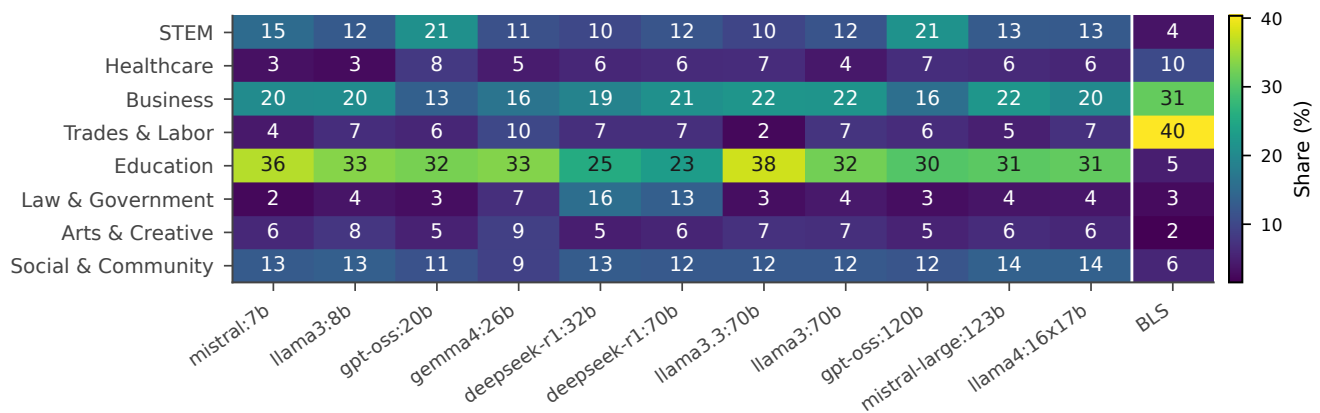


Figure 8: Occupational-category visibility by individual model (sorted by parameter count). Cross-family spread is real but small, and all 11 models share the same directional skew.

Similarity to Open-Web Career Advice

Prior work identifies web-text composition as one plausible source of occupational skew, alongside fine-tuning, alignment, and human feedback (Bender et al. 2021; Gallegos et al. 2024). The LLM distribution is closest in Jensen–Shannon divergence to “General Advice / Blogs” (JSD 0.021), markedly farther from user-generated video and personal-finance content, and far from the BLS structure, consistent with a model that reflects open-web written career advice rather than the labor market. We read this as evidence *consistent with* web-text mirroring rather than proof of a specific causal mechanism, since the source-type labels come from a single coder, the open web only proxies any

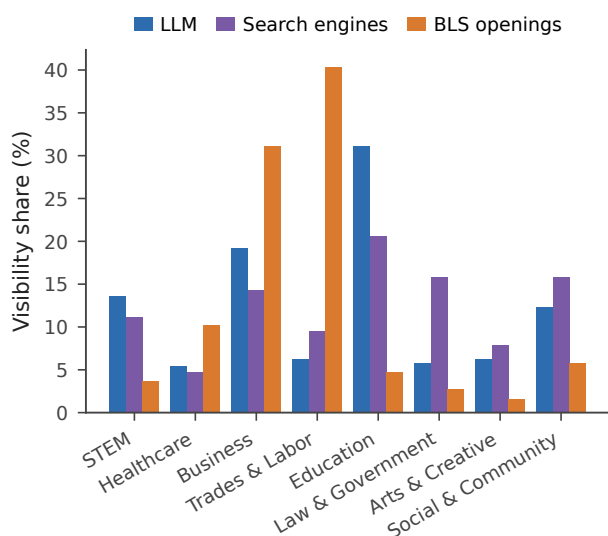


Figure 9: Visibility in search-engine results vs. BLS-projected job openings (2024–2034), applying the same keyword scheme used for LLM analysis.

particular training mixture, and the pattern could equally arise through instruction tuning, alignment, or shared web culture rather than pretraining alone.

Socioeconomic Steering

Models do respond meaningfully to context, which is often desirable, as when a creative individual is pointed toward art-focused careers. The danger is that the same sensitivity, applied to a one-sentence socioeconomic cue, behaves as a mirror rather than an advisor (Zajko 2022). We frame this as the model’s response to a cue rather than its treatment of any real socioeconomic group, but the cue-sensitivity is itself the worry, since a single sentence about financial constraint shifts the credentialing tier the model surfaces when it discusses credentials at all. Bachelor’s-or-above is 8.3% of Low-Income tier mentions against 73.4% for Wealthy/Privileged, and because the sectors surfaced to

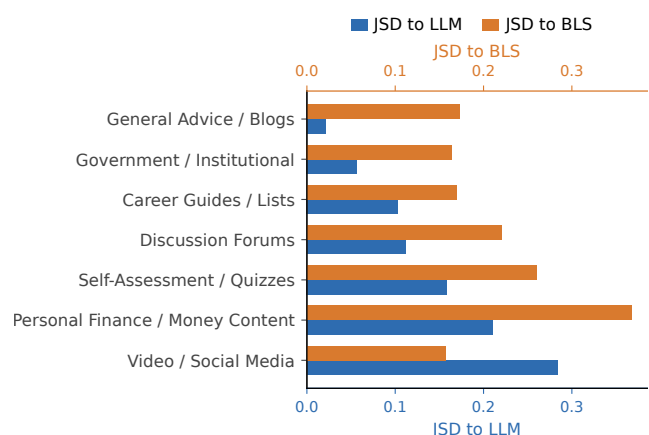


Figure 10: Jensen–Shannon divergence between each search-result source type and the LLM aggregate (blue, bottom axis) and the BLS baseline (orange, top axis). The LLM aggregate is closest to “General Advice / Blogs.”

the two groups imply almost identical wages, essentially the entire 50% education-implied wage gap is carried by the credentialing tier recommended within the same fields. The model does not withhold bachelor's pathways from Low-Income users so much as route them to lower-credential, lower-wage entry points into those fields, reproducing the social logic of constrained opportunity (Schoon and Polek 2011).

Design Implications

These patterns point to two simple design changes. First, because users cannot see what the model leaves out and a fluent answer sounds more authoritative than it should (Bender et al. 2021), career-advising tools should say plainly that their suggestions reflect patterns in training and alignment data rather than a labor-market census, and should point users to BLS data or a human advisor for a fuller picture. Second, the credentialing skew suggests a potential mitigation: whenever the model recommends a Bachelor's-or-above path, it should also list the sub-degree routes (apprenticeships, certificates, associate's programs) into the same field. The goal is not to dictate any particular recommendation but to make what is being left out visible to the user.

Limitations and Future Work

Keyword matching cannot capture nuanced or hybrid recommendations, and the Education category in particular is inflated because the token "education" appears in generic upskilling advice, which is why we flag rather than rely on Education figures. The SOC-to-category mapping also involves boundary calls, such as whether Personal Care & Service (SOC 39) belongs to Social & Community or Trades & Labor, that move the affected shares by a few points without altering the central Trades & Labor finding. The education-tier results rest on two approximate signals, the BLS within-category education distribution and the model's explicit tier mentions (present in only 12% of responses), which we report together so that each corroborates the other.

Several factors bound generalizability. The profiles are synthetic and briefly stated, so profile-level effects may differ for real users with richer self-presentation. The baseline is the US BLS projection and all interaction is in English, so the gaps could change under other labor markets, languages, or taxonomies, and the 11 largely open-weight models are a point-in-time sample that proprietary assistants may not match, though the cross-family and cross-temperature consistency argues against idiosyncrasy. The search-engine comparison is a contextual reference rather than a neutral ground truth. Engines personalize results while our queries carried no session state, locale, or recorded date. The comparison scans only result titles and URLs rather than full page text, and the source-type labels come from a single coder. Because the search snapshot was collected contemporaneously with the models, some results may themselves be LLM-influenced, so isolating pre-LLM web text, which we leave to future work, would require an archival corpus we do not have.

The most direct extension is to replace keyword matching with embedding-based similarity, scoring visibility by

cosine distance between response text and occupation descriptions, which would capture paraphrase and dissolve the Education inflation, at the cost of transparency and a risk of circularity if an LLM-derived encoder audits LLM output. Reporting both metrics with human-validated relevance judgments would test whether the documented skews are semantic rather than lexical.

Two empirical extensions follow directly. The profile-level effects should be validated with real users issuing naturalistic, multi-turn queries rather than synthetic personas, testing whether the credentialing asymmetry survives richer self-presentation. The mitigation we propose, surfacing sub-degree alternatives whenever the model recommends a Bachelor's-or-above pathway, can also be evaluated directly for whether it widens the visible opportunity set without degrading perceived helpfulness.

Conclusion

We characterized life-path visibility in LLM-generated guidance across 165,000 responses spanning 100 profiles, 15 prompts, 11 models, two role framings, and five temperatures, using a keyword scheme derived automatically from BLS occupation titles. The central gap is clear. Trades & Labor accounts for roughly 40% of projected job openings but only 6% of visibility, the largest absolute gap in the data, while Business is under-visible (19% vs. 31%) and Education and STEM are surfaced far above their labor-market shares.

Re-weighting by entry-level education exposes a credentialing skew, with implied Bachelor's-or-above openings at 44.8% against a BLS share of 20.6%, and this is where the socioeconomic harm concentrates. When responses discuss an education tier (28.9% of Low-Income vs. 4.1% of Wealthy/Privileged responses), Low-Income profiles receive far less Bachelor's-or-above guidance (8.3% vs. 73.4% of mentions), carrying a 50% education-implied wage gap. Unconditionally, the two groups are comparable (4.2% vs. 3.1%), so the harm lies in the *direction* of credential advice, not overall exposure. The aggregate distribution most resembles open-web "General Advice / Blogs" content and lies far from the labor market, consistent with web-text mirroring. Users reasonably expect a representative view of their options, yet these distributional patterns quietly narrow the opportunity landscape in ways they cannot detect.

A Reproducibility

All code and data needed to reproduce the results in this paper are available at the link below.

<https://academictorrents.com/details/a77baccadceb644bed72cdbfb7cca22a08b493f7>

References

Anderson, B. R.; Shah, J. H.; and Kreminski, M. 2024. Homogenization Effects of Large Language Models on Human Creative Ideation. In *Proceedings of the 16th Conference on Creativity & Cognition (C&C)*, 413–425.

- Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3): 337–351.
- Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.
- Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 610–623.
- Buçinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 188:1–188:21.
- Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics Derived Automatically from Language Corpora Contain Human-Like Biases. *Science*, 356(6334): 183–186.
- Deshpande, A.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; and Narasimhan, K. 2023. Toxicity in ChatGPT: Analyzing Persona-Assigned Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1236–1270.
- Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.
- Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3): 1097–1179.
- Government Digital Service. 2026. AI Insights: Large Language Models (LLMs) Bias. <https://www.gov.uk/government/publications/ai-insights/ai-insights-large-language-models-llms-bias-html>. Accessed: 2026-08-24.
- Kantharuban, A.; Milbauer, J.; Sap, M.; Strubell, E.; and Neubig, G. 2025. Stereotype or Personalization? User Identity Biases Chatbot Recommendations. In *Findings of the Association for Computational Linguistics: ACL 2025*, 24418–24436.
- Lent, R. W.; Brown, S. D.; and Hackett, G. 1994. Toward a Unifying Social Cognitive Theory of Career and Academic Interest, Choice, and Performance. *Journal of Vocational Behavior*, 45(1): 79–122.
- Logg, J. M.; Minson, J. A.; and Moore, D. A. 2019. Algorithm Appreciation: People Prefer Algorithmic to Human Judgment. *Organizational Behavior and Human Decision Processes*, 151: 90–103.
- Nghiem, H.; Prindle, J.; Zhao, J.; and Daumé III, H. 2024. “You Gotta be a Doctor, Lin”: An Investigation of Name-Based Bias of Large Language Models in Employment Recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7268–7287.
- Noble, S. U. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.
- Richardson, R.; Schultz, J. M.; and Crawford, K. 2019. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. *New York University Law Review Online*, 94: 192–233.
- Salinas, A.; Shah, P. V.; Huang, Y.; McCormack, R.; and Morstatter, F. 2023. The Unequal Opportunities of Large Language Models: Examining Demographic Biases in Job Recommendations by ChatGPT and LLaMA. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*, 34:1–34:15.
- Schoon, I.; and Polek, E. 2011. Teenage Career Aspirations and Adult Career Attainment: The Role of Gender, Social Background, and General Cognitive Ability. *International Journal of Behavioral Development*, 35(3): 210–217.
- Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 59–68.
- Shi, Y.; Yu, K.; Dong, Y.; and Chen, F. 2026. Large Language Models in Education: A Systematic Review of Empirical Applications, Benefits, and Challenges. *Computers and Education: Artificial Intelligence*, 10: 100529.
- U.S. Bureau of Labor Statistics. 2025. Employment Projections: 2024–34. <https://www.bls.gov/emp/>. Accessed: 2026-05-20.
- Vasconcelos, H.; Jörke, M.; Grunde-McLaughlin, M.; Gerstenberg, T.; Bernstein, M. S.; and Krishna, R. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 129:1–129:38.
- Wan, Y.; Pu, G.; Sun, J.; Garimella, A.; Chang, K.-W.; and Peng, N. 2023. “Kelly is a Warm Person, Joseph is a Role Model”: Gender Biases in LLM-Generated Reference Letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3730–3748.
- Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; et al. 2021. Ethical and Social Risks of Harm from Language Models. *arXiv preprint arXiv:2112.04359*.
- Wilson, K.; and Caliskan, A. 2024. Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 1578–1590.
- Zajko, M. 2022. Artificial Intelligence, Algorithms, and Social Inequality: Sociological Contributions to Contemporary Debates. *Sociology Compass*, 16(3): e12962.