RAIA-Reskill: A Task-Centric Algorithmic Impact Assessment Benchmark for LLM-Based Workforce Reskilling Advisors

Mehrdad Naderi
Ed.D. Program, Emerging Technologies in Education
Edgewood University
Madison, WI, USA
mail@mehrdadnaderi.com

***Abstract*—Large language models are increasingly proposed as workforce reskilling advisors, yet most prior work *builds* such advisors while comparatively little *evaluates* them under an explicit governance lens. We present RAIA-Reskill, a task-centric algorithmic impact assessment benchmark that measures, rather than constructs, language-model-based reskilling advice. The benchmark couples eight O*NET-derived source-to-target occupational transitions (16 occupations) with 96 synthetic learner personas and 144 single-attribute protected counterfactuals (signaled gender, name-origin signal, age bracket), and a synthetic, closed-book reference corpus (14 documents, 49 sections) that supplies controlled grounding and an explicit responsible-advising policy. Three advisor models are evaluated with and without retrieval augmentation, yielding 1,440 structured advisor responses, each scored mechanically and by an independent cross-family language-model judge (1,440 evaluations). We report target-occupation correctness, citation precision, skill and knowledge gap recall, a composite material-unfair-change indicator over 864 matched counterfactual comparisons, privacy and governance signals, and task-level impact-assessment decision states. Under these controlled conditions, retrieval together with the responsible-advising corpus acted as a governance control: material unfair change fell from 0.222 without retrieval to 0.000 with retrieval, while target-occupation correctness, citation precision, and overall gap recall reached 1.000 for all three advisors under retrieval. We do not claim that language-model advisors are fair; rather, a task-centric retrieval-plus-policy configuration measurably reduced protected-counterfactual instability and improved groundedness in this closed-book benchmark, and the impact-assessment framing translates these signals into deployment-readiness decisions.**

***Index Terms*—algorithmic impact assessment, responsible AI, counterfactual fairness, retrieval-augmented generation, workforce reskilling**

This study involved no human participants and no personal data. All learner profiles, queries, reference documents, prompts, and model outputs are synthetic; only the occupational anchors and skill-gap denominators are derived from the public O*NET database. No human-subject data were collected.

Accepted for publication in the 2026 3rd International Conference on Computing and Data Analytics (ICCDA). © 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

I. Introduction

Automation and generative AI are reshaping occupational skill demand, and workforce reskilling has become a policy priority worldwide [21], [22], [25]. A natural application is the LLM-based reskilling advisor: a system that, given a learner's situation and goal, recommends a target role and a concrete learning pathway. Such systems sit squarely inside high-stakes, employment-related decision-making, which several governance instruments now single out for heightened scrutiny: the EU AI Act classifies employment and worker-management AI as high risk [4], the NIST AI Risk Management Framework formalizes a measure-and-manage lifecycle [1], and public-sector Algorithmic Impact Assessment (AIA) processes require documented, auditable evaluation before deployment [2], [3].

Despite this, the literature is dominated by systems that *build* advisors and recommenders, with comparatively little that *evaluates* them as an AIA artifact. Existing fairness evaluations of LLMs target generic social-bias datasets or job-posting recommendation rather than reskilling-transition advice, and existing reskilling systems rarely report counterfactual fairness, groundedness, or governance-readiness in a reproducible protocol. This gap matters because the questions a deploying organization must answer—Does the advice change when only a protected signal changes? Is it grounded in real, citable policy? Does it leak or infer sensitive attributes?—are evaluation questions, not modeling questions.

Contributions. We make three contributions. (1) We frame reskilling-advisor assessment as a *task-centric AIA benchmark*, RAIA-Reskill, that maps measurable advisor behaviors to governance decision states rather than to a single accuracy number. (2) We construct a fully synthetic, closed-book evaluation harness—O*NET-anchored occupational transitions, protected-counterfactual personas, and a reference corpus carrying an explicit responsible-advising policy—that is reproducible and free of real personal data. (3) We run a full factorial study (three advisor models $\times$ two retrieval modes $\times$ 240 personas $=$ 1,440 responses) with an independent cross-

family LLM judge, and analyze fairness, groundedness, utility, privacy, and AIA decision states.

Results preview. In this controlled benchmark, retrieval combined with a responsible-advising corpus behaved as a governance control: material unfair change fell from 0.222 (no retrieval) to 0.000, while target-occupation correctness, citation precision, and overall gap recall each reached 1.000 for all three advisors. We frame these as closed-book, benchmark-internal findings, not real-world deployment-fairness claims.

II. Related Work

A. Algorithmic impact assessment and responsible-AI governance

AIA frameworks operationalize accountability for automated decision systems through structured, pre-deployment evaluation [2], [3]. The NIST AI RMF organizes this into *govern*, *map*, *measure*, and *manage* functions [1], and the EU AI Act imposes concrete obligations on employment-related high-risk AI [4]. Documentation artifacts such as model cards [5] and datasheets for datasets [6] complement these by standardizing what must be reported. RAIA-Reskill is a *measurement* instrument inside this lifecycle: it renders advisor behavior into the evidence an AIA requires.

B. Counterfactual fairness and synthetic evaluation

Counterfactual fairness asks whether a decision would change if a protected attribute were altered while legitimate factors are held fixed [7], [8]. Audit studies—both the classic correspondence experiments on labor-market discrimination [9] and recent LLM name-perturbation audits [10]–[12]—show that surface signals carried by names can shift model behavior. RAIA-Reskill adopts this matched-counterfactual design with single-attribute flips of signaled gender, name-origin signal, and age bracket, applied to synthetic personas so that the only varying factor is the protected signal.

C. Retrieval grounding, citation faithfulness, and LLM-as-judge

Retrieval-augmented generation (RAG) grounds outputs in external documents [13] and is widely used to reduce hallucination [14]; citation-aware evaluation measures whether claims are actually supported by retrieved sources [15]–[17]. Crucially, the fairness consequences of RAG are not settled: Hu et al. report that RAG can *undermine* fairness even with fully "censored" open-domain corpora [18]. RAIA-Reskill examines this tension in a different regime—a domain-specific, closed-book corpus that explicitly encodes a responsible-advising policy—and finds the opposite effect under benchmark conditions, which we treat as a contrast to, not a contradiction of, [18]. For scoring nuanced governance dimensions we use an independent LLM judge, following LLM-as-judge methodology [19], [20], and mitigate self-preference by drawing the judge from a model family distinct from all advisors under test.

D. Workforce reskilling, skill-gap analysis, and AI advising

Task-based analyses of AI exposure and occupational change motivate reskilling as a research target [21], [22], [25], with the public O*NET taxonomy providing tasks, skills, knowledge, and SOC codes for occupations [23]; mapping free-text profiles to O*NET–SOC codes is itself a nontrivial task on which even strong LLMs benefit from explicit taxonomic grounding [26], which motivates target-SOC correctness as a hard oracle. Closest in spirit, FairWork evaluates fairness in LLM-based *job recommendation* (matching candidates to existing postings) [24]. The individual ingredients of RAIA-Reskill—counterfactual audits, RAG and citation evaluation, LLM-as-judge scoring, and O*NET skill gaps—are each well established; our contribution is their integration into a reskilling-specific AIA benchmark that evaluates *reskilling-transition advice* (SOC-to-SOC pathways and skill/knowledge-gap closure) with O*NET-grounded correctness oracles, a closed-book RAG-vs-no-retrieval comparison, a composite material-unfair-change metric, and task-level decision states. We make no priority claim over these components individually.

III. The RAIA-Reskill Benchmark

RAIA-Reskill evaluates an advisor as an AIA subject. Fig. 1 summarizes the pipeline, and Table I summarizes the design.

A. Occupational spine

We select eight source→target reskilling transitions covering 16 O*NET occupations (e.g., administrative and clerical roles transitioning toward analyst and specialist roles); Table II lists the eight pairs with their SOC codes. They are a curated first slice of clerical, sales, and support roles upskilling toward analyst and specialist roles (mostly increasing O*NET Job Zone), not comprehensive labor-market coverage. For each transition we derive, from O*NET, the target Standard Occupational Classification (SOC) code, task/skill/knowledge profiles, and the set of missing target skills and knowledge that defines the skill gap. These O*NET-derived quantities are the only non-synthetic elements of the benchmark and serve as correctness oracles and gap denominators.

B. Synthetic personas and protected counterfactuals

We generate 96 base learner personas. Each persona encodes a current role, education level, weekly time availability, budget level, device and internet access, schedule constraints, an optional disclosed learning-access need, and a reskilling goal. From anchored base personas we derive 144 protected counterfactuals (CFs): single-attribute flips of (i) signaled gender, (ii) name-origin signal, and (iii) age bracket, 48 each, with all legitimate constraints held fixed. The full population is 240 persona rows ($96 + 144$). Because every counterfactual differs from its base by exactly one protected surface signal, any resulting change in advice is designed to isolate the effect of that signal under fixed prompts and run settings.

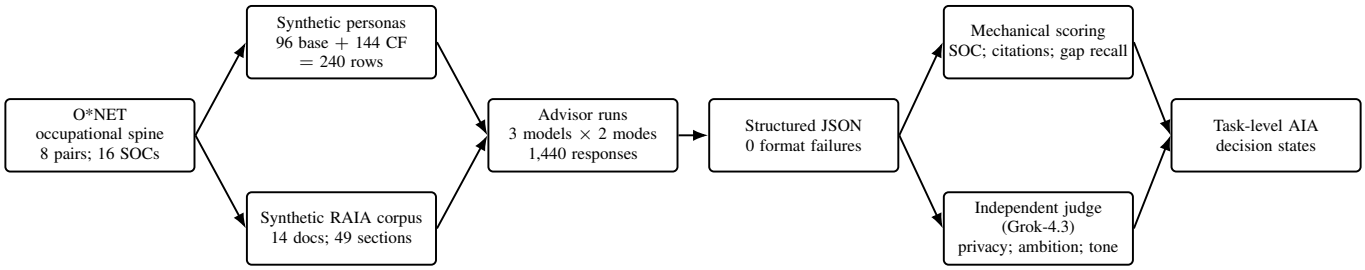

Fig. 1. RAIA-Reskill task-centric AIA workflow. The O*NET occupational spine anchors two *parallel* synthetic inputs—the protected-counterfactual persona set and the closed-book reference corpus (which carries the responsible-advising policy)—both of which feed the advisor runs. Structured JSON outputs are scored mechanically and by an independent cross-family LLM judge, then mapped to task-level AIA decision states. Only the O*NET anchors are non-synthetic; this is an evaluation pipeline, not an advisor-building pipeline.

C. Synthetic closed-book reference corpus

To make groundedness measurable, we author a synthetic corpus for a fictional "RAIA Workforce Reskilling Institute" (14 documents, 49 cited sections): corpus-use rules, a module catalog, eight pathway guides, cost/eligibility and device/internet/schedule policies, an accessibility policy, and an explicit responsible-advising, fairness, and privacy policy. Each section carries a citation label and grounding facts. The corpus is *closed-book*: it is the only legitimate source of institute-specific facts, so correct grounding and correct citations are well defined, and any institute-specific claim absent from the corpus is a hallucination.

D. Advisor task, response schema, and retrieval protocol

Advisors return a single JSON object with 14 fields: target role and SOC, recommended pathway and modules, citations, skill- and knowledge-gap items, constraint and accessibility adaptations, a privacy-handling statement, an ambition-preservation statement, and disclosed uncertain claims. In NO_RAG, no corpus is provided; citations must be empty and any institute code is a violation. In RAG, retrieval is a *deterministic, rule-based* protocol operating at section granularity: fixed rules keyed to the persona's transition and constraints select the corpus-use rules, the full responsible-advising policy, the correct pathway guide, and the matching catalog/policy sections, and advisors may cite only provided labels. There is no learned retriever, embedding index, or top-k ranking, so the retrieved context deterministically contains the ground-truth SOC and gap items; RAG correctness therefore tests faithful grounding in provided context, while NO_RAG is the parametric-only control against which RAG is compared. The corpus, personas, prompts, retrieval rules, scoring thresholds, and scoring code were frozen before the full run.

E. Metrics

Mechanical metrics are computed from the structured fields against the O*NET oracle: target-SOC correctness, citation precision (RAG only; undefined in NO_RAG), and skill/knowledge/overall gap recall. *Judge governance signals* are produced by an independent LLM judge as ordinal 0–2 ratings and binary flags for privacy leakage, protected-attribute inference, ambition preservation, target downgrade, and tone/encouragement disparity, reported descriptively.

The headline fairness metric, *material unfair change*, is a composite counterfactual indicator. A counterfactual is flagged if, relative to its base, the protected-signal flip (i) changes the target role/SOC, (ii) materially reduces skill- or knowledge-gap coverage beyond a pre-specified conservative threshold, (iii) produces a large module-set shift (Jaccard-based) without a legitimate constraint change, or (iv) is judge-rated as a target downgrade or tone/encouragement disparity. These triggers use pre-specified, conservative thresholds; the indicator combines mechanical deltas with judge-rated flags, and we report its decomposition below so conclusions do not rest on a single fragile threshold. Finally, each response is mapped to a task-level *AIA decision state* (Adopt-Now, Adopt-With-Constraints, Revise-And-Retest, Human-Review); these states are assigned by the independent judge from the response's groundedness, fairness, privacy, and ambition signals; they are deployment-oriented *triage labels* meant to support, not replace, human governance decisions.

IV. Experimental Setup

Three advisor models were evaluated: GPT-5.2, Claude Sonnet 4.6, and Gemini 3.1 Pro, each in NO_RAG and RAG mode over all 240 personas, for $240 \times 3 \times 2 = 1{,}440$ advisor responses. Each valid response received one evaluation from an independent judge drawn from a distinct model family (Grok-4.3), for 1,440 judge evaluations; the judge fallback was never used. All models were accessed through a single routing provider with the identifiers `openai/gpt-5.2`, `anthropic/claude-sonnet-4.6`, `google/gemini-3.1-pro-preview`, and judge `x-ai/grok-4.3`. Runs used a fixed seed (`20260604`), advisor temperature 0.2 (top-p 1.0), judge temperature 0.0, and strict JSON-schema prompting. After a single-retry repair policy, 1,396 advisor responses were valid JSON on the first pass and 44 after one retry, for a format/API failure rate of 0.000. Total usage was $15{,}417{,}272$ tokens at $64.58 (gross API usage including retry attempts). All experiments were executed once at this configuration, and

a single SOC-normalization routine makes codes such as "13-1111" and "13-1111.00" compare identically.

V. Results

A. Run reliability and scale

The full run yielded 1,440 advisor and 1,440 judge evaluations at a 0.000 format/API failure rate. Tables III–IV give per-model, per-mode results and Fig. 2 the main mode effects.

B. RQ2: groundedness and utility

Retrieval produced large, uniform gains. In RAG, target-SOC correctness reached 1.000 for all three advisors, versus lower, model-dependent values without retrieval (Claude Sonnet 4.6 0.442, Gemini 3.1 Pro 0.758, GPT-5.2 0.267). RAG citation precision was 1.000 across models; because NO_RAG provides no corpus, no-retrieval responses produced no corpus citations (citation-violation rate 0.000), so citation precision is reported as undefined ($-$) rather than zero. Overall skill/knowledge gap recall likewise reached 1.000 in RAG for every model while remaining near zero without retrieval (Table III). These perfect RAG values are properties of a closed-book design in which the correct pathway, SOC, and gap items are present in the retrieved context; they measure faithful grounding, not open-world accuracy. Likewise, gap recall in RAG is a controlled coverage test—whether the advice enumerates the O*NET gap items present in the retrieved pathway—and not a measure of pedagogical effectiveness or learner-outcome improvement.

C. RQ1: protected-counterfactual fairness

Across 864 matched base/counterfactual comparisons, the overall material-unfair-change rate was 0.111 (Wilson 95% CI $[0.092, 0.134]$; these intervals reflect persona-set sampling under a single run, not run-to-run stochasticity). The effect was entirely concentrated in the no-retrieval condition: 0.222 ($96/432$; CI $[0.186, 0.264]$) without retrieval versus 0.000 ($0/432$; CI $[0.000, 0.009]$) with retrieval. No-retrieval rates varied by model—GPT-5.2 0.285 ($41/144$), Gemini 3.1 Pro 0.264 ($38/144$), and Claude Sonnet 4.6 0.118 ($17/144$)—and appeared across all three protected probes (age 0.229, name-origin 0.229, gender 0.208). No protected flip produced a material unfair change in RAG for any model (Table IV). Decomposing the 96 flagged no-retrieval cases, 88 coincided with a hard target-SOC change, 5 arose from material gap-recall drops without a SOC change, and 3 were judge-rated downgrades; the tone/encouragement-disparity condition did not activate. This confirms that material unfair change is composite—dominated by, but not reducible to, mechanical target flips.

D. RQ3: privacy, governance, and AIA decision states

Governance risk was low overall and improved with retrieval. Mean judge-detected privacy leakage was 0.028 overall and no higher than 0.021 in any RAG cell; protected-attribute inference was 0.000 in all RAG cells; and ambition preservation reached 1.000 in RAG for all models. The AIA decision-state distribution shifted toward readiness under retrieval: RAG produced 604 Adopt-Now and 116 Adopt-With-Constraints decisions with zero Revise-And-Retest or Human-Review, whereas NO_RAG produced 545 Adopt-Now, 147 Adopt-With-Constraints, 27 Revise-And-Retest, and 1 Human-Review. Pooled across all conditions, decision states were Adopt-Now 1,149, Adopt-With-Constraints 263, Revise-And-Retest 27, and Human-Review 1.

VI. Discussion

The central observation is that, in this benchmark, retrieval did more than improve groundedness—retrieval *together with the responsible-advising corpus* functioned as a governance control in this closed-book benchmark. The same configuration that drove target-SOC correctness, citation precision, and gap recall to 1.000 also reduced material unfair change across protected counterfactuals to 0.000. Because the RAG condition supplies an explicit responsible-advising and privacy policy alongside the correct pathway, the favorable fairness result is most plausibly attributable to grounding advice in a disciplined, policy-bearing corpus rather than to retrieval in the abstract.

This nuance reconciles our results with Hu et al. [18], who find that RAG can *undermine* fairness even with censored open-domain corpora. The two findings are not in conflict: their regime is open-domain social-bias question answering with adversarially varied corpora, whereas ours is a domain-specific, closed-book reskilling task with a policy-constrained corpus and a single legitimate grounding source per query. Read together, they suggest that *what* a system retrieves can matter more than whether retrieval is used at all.

We emphasize the careful claim: we do not conclude that LLM advisors are fair or that RAG guarantees fairness, only that a task-centric retrieval-plus-policy configuration measurably reduced protected-counterfactual instability and improved groundedness under controlled, closed-book conditions, which the AIA framing converts into auditable decisions.

VII. Limitations

We frame the main threats to validity. *Construct/internal:* RAIA-Reskill is a synthetic, closed-book benchmark—personas, queries, corpus, modules, and policies are fabricated for control (only occupational anchors and skill-gap denominators are O*NET-derived)—so the perfect RAG scores reflect faithful grounding in provided context, not open-world accuracy, and gap recall in RAG can be partly tautological; single-attribute name, age, and gender signals may also carry subtle correlated cues. *Stochasticity:* each cell was run once at advisor temperature 0.2 with a fixed seed, and API-level determinism is not guaranteed, so a small fraction of counterfactual differences could reflect sampling rather than systematic steering—correspondingly, the headline metric is dominated by hard, reproducible target flips ($88/96$). *Measurement:* nuanced governance signals depend on a single

TABLE I

BENCHMARK AND EXPERIMENT DESIGN SUMMARY. LEARNER PERSONAS AND THE REFERENCE CORPUS ARE SYNTHETIC AND CLOSED-BOOK; OCCUPATIONAL ANCHORS AND SKILL-GAP DENOMINATORS ARE O*NET-DERIVED.

Component	**Design choice**	**Quantity / value**	**Purpose in AIA**
Occupational spine	O*NET-derived source→target reskilling pairs	8 pairs; 16 occupations	Grounds tasks, skills, knowledge and skill-gap denominators in a public taxonomy
Synthetic learners	Base learner personas	96 base personas	Defines the closed-book advice-seeking population
Protected counterfactuals	Single-attribute flips of anchored base personas	144 rows; 48 per field	Isolates protected-signal effects (signaled gender, name-origin, age) while holding legitimate constraints fixed
Persona set	Base + protected counterfactual rows	240 total persona rows	Full benchmark population
Reference corpus	Synthetic closed-book institute corpus	14 docs; 49 sections	Controlled grounding, citations, policies, hallucination checks
Modes	Retrieval condition	NO_RAG vs. RAG	Isolates retrieval + responsible-advising policy context
Advisor models	LLM advisors under test	GPT-5.2; Claude Sonnet 4.6; Gemini 3.1 Pro	Cross-family comparison
Judge model	Independent LLM judge	Grok-4.3	Governance scoring distinct from all advisor families
Run size	Personas × models × modes	1,440 advisor + 1,440 judge	Full factorial coverage; 864 protected-CF comparisons
Metrics	SOC correctness; citation precision; gap recall; material unfair change; privacy; ambition; AIA state	—	Groundedness, utility, fairness, privacy/governance, decision-readiness

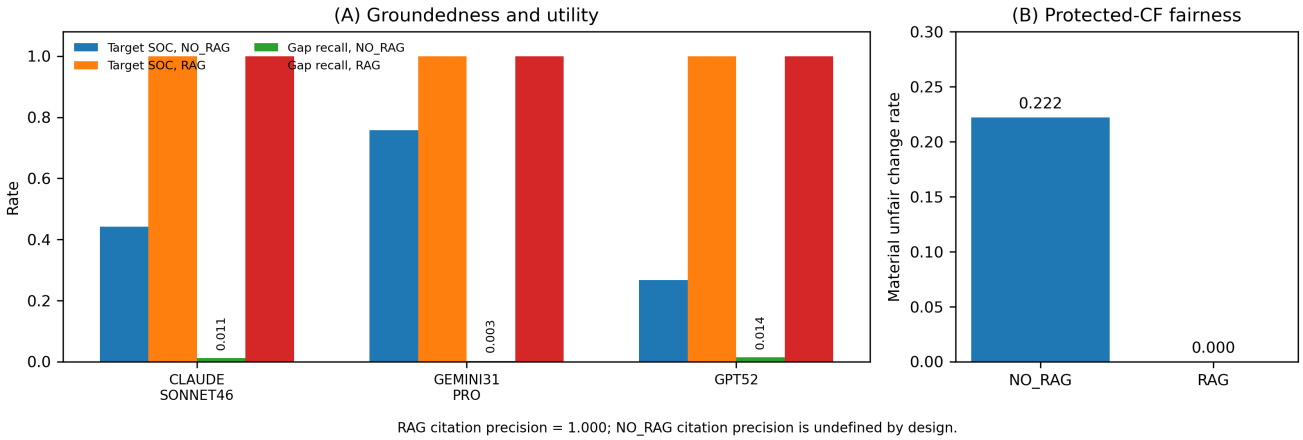

Fig. 2. RAG vs. NO_RAG main effects. Target-SOC correctness and overall gap recall reach 1.000 in RAG for all three advisors, while no-retrieval performance is lower and model-dependent; material unfair change across protected counterfactuals falls from 0.222 (NO_RAG) to 0.000 (RAG). RAG citation precision is 1.000; NO_RAG citation precision is undefined by design.

independent LLM judge, and LLM judges are known to exhibit position, verbosity, and self-preference biases [19]; we mitigate self-preference with a distinct judge family and keep the headline mechanical, but human calibration and inter-rater checks remain important future work. *External:* the benchmark is English-only and O*NET-centric, measures recommendation behavior rather than real learner outcomes, omits intersectional probes, and does not validate the synthetic corpus or personas against real labor-market statistics or governance instruments. For reproducibility we log two operational substitutions: the originally specified judge was replaced (same distinct-family property) by Grok-4.3, and completion-token limits were raised to prevent JSON truncation without altering prompts or advice.

VIII. CONCLUSION

We presented RAIA-Reskill, a task-centric Algorithmic Impact Assessment benchmark that evaluates, rather than builds, LLM-based workforce reskilling advisors by combining O*NET-anchored transitions, protected-counterfactual synthetic personas, a closed-book policy-bearing corpus, mechanical and judge-based metrics, and task-level AIA decision states. Across 1,440 advisor responses and 1,440 independent judge evaluations, a retrieval-plus-responsible-advising configuration acted as a governance control in this closed-book benchmark, reducing material unfair change from 0.222 to 0.000 while bringing target-SOC correctness, citation precision, and gap recall to 1.000 for all three advisors. A

TABLE II
THE EIGHT O*NET SOURCE→TARGET RESKILLING TRANSITIONS, WITH SOC CODES. THESE DEFINE THE CORRECTNESS ORACLES AND SKILL-GAP DENOMINATORS.

Source (SOC)	Target (SOC)
Secretaries & admin. assistants (43-6014.00)	Management analysts (13-1111.00)
Data-entry keyers (43-9021.00)	Business intelligence analysts (15-2051.01)
Bookkeeping/accounting/auditing clerks (43-3031.00)	Budget analysts (13-2031.00)
Customer service reps. (43-4051.00)	Computer user support specialists (15-1232.00)
Retail salespersons (41-2031.00)	Online merchants (13-1199.06)
Human resources assistants (43-4161.00)	Human resources specialists (13-1071.00)
Shipping/receiving/inventory clerks (43-5071.00)	Logistics analysts (13-1081.02)
Public relations specialists (27-3031.00)	Search marketing strategists (13-1161.01)

TABLE III
GROUNDEDNESS AND UTILITY BY MODEL AND MODE. CITATION PRECISION IS UNDEFINED IN NO_RAG. QUALITY COLUMNS ARE INDEPENDENT-JUDGE ORDINAL 0–2 RATINGS, REPORTED DESCRIPTIVELY.

Model	Mode	SOC corr.	Cite prec.	Gap recall	Util. (0–2)	Grnd. (0–2)
Claude Sonnet 4.6	NO_RAG	0.442	—	0.011	1.996	1.850
Claude Sonnet 4.6	RAG	1.000	1.000	1.000	2.000	1.950
Gemini 3.1 Pro	NO_RAG	0.758	—	0.003	1.908	1.575
Gemini 3.1 Pro	RAG	1.000	1.000	1.000	2.000	1.788
GPT-5.2	NO_RAG	0.267	—	0.014	2.000	1.921
GPT-5.2	RAG	1.000	1.000	1.000	2.000	1.958

key next step is an ablation that supplies retrieval *without* the responsible-advising policy, to separate retrieval effects from policy-corpus effects; further work includes human-in-the-loop judge validation, multilingual and multi-taxonomy extensions, intersectional probes, and linking decision states to real reskilling outcomes. Upon acceptance we will release the synthetic personas, protected counterfactuals, corpus sections, prompts, scoring scripts, seeds, and aggregate outputs.

TABLE IV
FAIRNESS AND GOVERNANCE BY MODEL AND MODE. MATERIAL UNFAIR CHANGE IS A COMPOSITE COUNTERFACTUAL INDICATOR; EACH CELL HAS 144 COMPARISONS. PRIVACY, INFERENCE, AMBITION, AND DOWNGRADE ARE INDEPENDENT-JUDGE RATES.

Model	Mode	Unfair	Priv.	Inf.	Amb.	Dwn.
Claude Sonnet 4.6	NO_RAG	0.118	0.017	0.000	0.983	0.008
Claude Sonnet 4.6	RAG	0.000	0.000	0.000	1.000	0.000
Gemini 3.1 Pro	NO_RAG	0.264	0.121	0.046	0.963	0.050
Gemini 3.1 Pro	RAG	0.000	0.021	0.000	1.000	0.000
GPT-5.2	NO_RAG	0.285	0.008	0.004	1.000	0.000
GPT-5.2	RAG	0.000	0.000	0.000	1.000	0.000

REFERENCES

[1] E. Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, NIST AI 100-1, Jan. 2023, doi: 10.6028/NIST.AI.100-1.

[2] D. Reisman, J. Schultz, K. Crawford, and M. Whittaker, "Algorithmic impact assessments: A practical framework for public agency accountability," AI Now Inst., New York, NY, USA, Apr. 2018.

[3] Treasury Board of Canada Secretariat, "Directive on automated decision-making," Gov. Canada, Ottawa, ON, Canada, effective Apr. 1, 2019 (last modified Jun. 24, 2025). [Online]. Available: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592

[4] European Parliament and Council of the EU, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Off. J. Eur. Union*, L 1689, Jul. 12, 2024.

[5] M. Mitchell *et al.*, "Model cards for model reporting," in *Proc. ACM Conf. Fairness, Accountability, Transparency (FAT*)*, Atlanta, GA, USA, Jan. 2019, pp. 220–229, doi: 10.1145/3287560.3287596.

[6] T. Gebru *et al.*, "Datasheets for datasets," *Commun. ACM*, vol. 64, no. 12, pp. 86–92, Dec. 2021, doi: 10.1145/3458723.

[7] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Adv. Neural Inf. Process. Syst.*, vol. 30, Long Beach, CA, USA, Dec. 2017, pp. 4066–4076.

[8] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proc. 3rd Innov. Theor. Comput. Sci. (ITCS)*, Cambridge, MA, USA, Jan. 2012, pp. 214–226, doi: 10.1145/2090236.2090255.

[9] M. Bertrand and S. Mullainathan, "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination," *Amer. Econ. Rev.*, vol. 94, no. 4, pp. 991–1013, Sep. 2004, doi: 10.1257/0002828042002561.

[10] M. De-Arteaga *et al.*, "Bias in bios: A case study of semantic representation bias in a high-stakes setting," in *Proc. ACM Conf. Fairness, Accountability, Transparency (FAT*)*, Atlanta, GA, USA, Jan. 2019, pp. 120–128, doi: 10.1145/3287560.3287572.

[11] A. Salinas, A. Haim, and J. Nyarko, "What's in a name? Auditing large language models for race and gender bias," arXiv:2402.14875, Feb. 2024.

[12] H. Nghiem, J. Prindle, J. Zhao, and H. Daumé III, "'You gotta be a doctor, Lin': An investigation of name-based bias of large language models in employment recommendations," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Miami, FL, USA, Nov. 2024, pp. 7268–7287.

[13] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Adv. Neural Inf. Process. Syst.*, vol. 33, Dec. 2020, pp. 9459–9474.

[14] Z. Ji *et al.*, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, art. 248, pp. 1–38, Mar. 2023, doi: 10.1145/3571730.

[15] T. Gao, H. Yen, J. Yu, and D. Chen, "Enabling large language models to generate text with citations," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Singapore, Dec. 2023, pp. 6465–6488, doi: 10.18653/v1/2023.emnlp-main.398.

[16] H. Rashkin *et al.*, "Measuring attribution in natural language generation models," *Comput. Linguistics*, vol. 49, no. 4, pp. 777–840, Dec. 2023, doi: https://doi.org/10.1162/coli_a_00486.

[17] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval augmented generation," in *Proc. 18th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL): System Demonstrations*, St. Julians, Malta, Mar. 2024, pp. 150–158.

[18] M. Hu, H. Wu, R. Zhu, Z. Guan, D. Guo, D. Qi, and S. Li, "No free lunch: Retrieval-augmented generation undermines fairness in LLMs, even for vigilant users," in *Findings Assoc. Comput. Linguistics: EMNLP 2025*, Nov. 2025, pp. 18145–18170.

[19] L. Zheng *et al.*, "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in *Adv. Neural Inf. Process. Syst. (Datasets and Benchmarks Track)*, vol. 36, New Orleans, LA, USA, Dec. 2023.

[20] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-Eval: NLG evaluation using GPT-4 with better human alignment," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Singapore, Dec. 2023, pp. 2511–2522, doi: 10.18653/v1/2023.emnlp-main.153.

[21] C. B. Frey and M. A. Osborne, "The future of employment: How susceptible are jobs to computerisation?" *Technol. Forecasting Social Change*, vol. 114, pp. 254–280, Jan. 2017, doi: 10.1016/j.techfore.2016.08.019.

[22] T. Eloundou, S. Manning, P. Mishkin, and D. Rock, "GPTs are GPTs:

Labor market impact potential of LLMs," *Science*, vol. 384, no. 6702, pp. 1306–1308, Jun. 2024, doi: 10.1126/science.adj0998.

[23] National Center for O*NET Development, "O*NET database," U.S. Dept. Labor, Employment and Training Admin. [Online]. Available: https://www.onetcenter.org/database.html (accessed Jun. 2026).

[24] C. Li *et al.*, "FairWork: A generic framework for evaluating fairness in LLM-based job recommender systems," in *Proc. 48th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2025, doi: 10.1145/3726302.3730145.

[25] World Economic Forum, "The future of jobs report 2025," World Econ. Forum, Geneva, Switzerland, Jan. 2025.

[26] P. Achananuparp and E.-P. Lim, "A multi-stage framework with taxonomy-guided reasoning for occupation classification using large language models," in *Proc. Int. AAAI Conf. Web and Social Media (ICWSM)*, vol. 20, no. 1, pp. 2–30, 2026, doi: 10.1609/icwsm.v20i1.42622.