Using language models to measure the demand-based wage gap

Elodie Chervin
Department of Economics, University of Oxford

Abstract

Social scientists increasingly use Large Language Models (LLMs) to measure unobserved constructs from text. I apply this idea to pay equity, where the central object is not the traditional raw or residual gender wage gap, but a demand-based wage gap: the difference in pay between predominantly female and predominantly male jobs after accounting for the task demands that make jobs comparable in value. I develop a measurement procedure that uses LLMs to read 50,000 online job postings in Quebec and Ontario and assign job-value scores following the logic of pay equity committees. Rather than treating these scores as ground truth, I discipline them with the statutory rubric, compare them with human-labelled reference data, and check whether they are sensitive to gender-coded language. I then use the calibrated job-value measures to estimate the demand-based wage gap and simulate how alternative pay equity rules, audit rates, and compliance costs could change firm behaviour. The 1996 Quebec Pay Equity Act supplies the institutional scoring template; the empirical goal is measurement and simulation.

1 Introduction and institutional context

A central challenge in empirical economics is measuring important quantities that are not directly observed. Pay equity is a prime example. Policies such as the 1996 Quebec Pay Equity Act are built around the idea that predominantly female and predominantly male jobs should receive equal pay when they involve work of comparable value. In practice, value is determined by committees or employers evaluating job classes across statutory factor families: required qualifications, responsibilities, effort, and working conditions. These committee-evaluated job scores are usually confidential and unobserved outside audits, which makes it hard to measure the demand-based wage gap that pay equity rules are designed to target.

I address this measurement failure by scraping 50,000 online job postings from Quebec and Ontario and using LLMs to approximate the job-value information that would normally be produced by pay equity committees. The estimand is the demand-based wage gap: the wage difference between female-dominated and male-dominated jobs after adjusting for estimated job value, rather than the conventional wage gap after adjusting only for worker, firm, or occupation controls.

This measurement task is risky if LLM outputs are treated as objective scores. A model may inherit historical gendered valuations embedded in text, and small changes in wording can change the distribution of predicted scores. I therefore use a two-stage strategy. First, I score job postings with a fixed pay equity rubric and screen the scores for basic institutional consistency. Second, I calibrate the scores against human-labelled reference data using sample splitting. The calibrated measures then feed into wage-gap estimation and counterfactual simulations of possible policy effects.

2 Measuring job value from postings

Standard calibration approaches often start after the automated measure has already been produced. I instead build the measure around the institutional object I want to recover: the committee-style assessment of job value. The objective is not only to predict a small labelled set well, but to preserve the ranking and variation in task demands that determine comparable job value.

2.1 Rubric-based scoring

For each posting, the model reads the job description and assigns scores for the four statutory factor families: required qualifications, responsibilities, effort, and working conditions. The score is constructed in three steps.

Task extraction: I first extract the concrete task demands in the posting, such as supervision, required certification, exposure to hazards, physical effort, decision-making authority, and responsibility for people or budgets. This creates an intermediate job profile that can be compared across very different occupations.

Factor scoring: I map the extracted tasks to the four statutory factor families and retain both an ordinal factor score and a continuous score. The continuous score is used for wage-gap estimation, while the ordinal version is used when the simulation requires legally admissible pay equity categories.

Validation and scale adjustment: I compare the model scores with human-labelled examples from public-sector audits, collective agreements, and hand-coded benchmark jobs when a common scoring scale is available. These labelled examples anchor the scale and reveal where the model systematically over- or under-scores particular kinds of work.

2.2 Institutional consistency checks

The key threat is not random noise alone. The model may reproduce precisely the undervaluation that pay equity rules are meant to correct. I therefore impose three checks before using the scores in the wage-gap analysis.

Dominance and monotonicity: If one job clearly requires more certification, supervision, responsibility, or adverse working conditions than another, the corresponding factor score should not fall. Violations are flagged and either corrected through the scoring rule or excluded from the main specification.

Gender-word sensitivity: I create paired descriptions that hold tasks fixed while changing gender-coded wording. A valid job-value measure should respond to the tasks, not to whether the language sounds stereotypically masculine or feminine.

Theory-guided shrinkage: Following the logic of Fessler and Kasy (2019), I treat the legal and economic restrictions as soft priors. When a score is weakly supported by the text and violates a clear restriction, I shrink it toward the institutionally consistent range; when the text strongly supports an exception, I preserve the variation and flag it for validation. This keeps the measure disciplined without forcing every posting into a rigid template.

3 Demand-based wage-gap estimation

Even with these safeguards, the LLM score remains a noisy measure of true job value $Y_i^\dagger$. I therefore use human-labelled reference data to calibrate the LLM scores before estimating the demand-based wage gap. Let F_i indicate that job i belongs to a predominantly female job class. The target regression is

$$W_i = \delta F_i + \beta Y_i^\dagger + X_i^\top \gamma + u_i, \tag{1}$$

where δ is the demand-based wage gap: the wage difference associated with female job-class predominance after conditioning on estimated job value and other controls X_i. This differs from a traditional wage-gap decomposition because the key control is not only worker or firm composition, but the demand-based value of the work itself.

I use a prediction-powered correction to account for the fact that $Y_i^\dagger$ is observed only in a small validation sample. Let Y_i^* denote the human-labelled job-value score in the labelled sample $\mathcal{L}$, and let Y_i^{LLM} denote the LLM score available for both labelled and unlabelled postings. With sample splitting to avoid training leakage, the calibrated demand-gap estimator is

$$\hat{\delta}_{\text{calibrated}} = \hat{\delta}\big(Y_{\mathcal{U}}^{\text{LLM}}\big) + \left[\hat{\delta}\big(Y_{\mathcal{L}}^*\big) - \hat{\delta}\big(Y_{\mathcal{L}}^{\text{LLM}}\big)\right], \tag{2}$$

where $\hat{\delta}(Y)$ denotes the coefficient on F_i from the wage regression using job-value measure Y. This correction is valid under the usual validation-sample assumptions from prediction-powered inference: the labelled sample must identify measurement error for the target population, and uncertainty must account for the correction term. When multiple reference datasets are available, I use them as source-specific validation checks or as anchored labels when a common reference scale is available; I do not rely on cross-source bias differences without an anchoring assumption.

4 Counterfactual policy simulations

The calibrated job-value measures can be used to simulate possible effects of pay equity rules. The simulations ask how firms might respond under alternative audit probabilities, fines, comparison methods, and administrative costs.

Firms choose a compliance strategy $s \in \{\text{Ignore}, \text{Individual Method}, \text{Global Method}\}$ and reported job scores to minimise total expected costs:

$$\min_{s,\tilde{Y}_F} \sum_{j\in\{M,F\}} w_j(s)L_j + \text{AdminCost}(s) + C(\tilde{Y}_F, Y_F^\dagger) + \mathbb{E}[\Phi(s)] \tag{3}$$

where $w_M(s)L_M + w_F(s)L_F$ is the firm wage bill, $C(\tilde{Y}_F, Y_F^\dagger)$ is the cost of reporting a female-job score $\tilde{Y}_F$ below the calibrated job value $Y_F^\dagger$, and $\mathbb{E}[\Phi(s)]$ is the expected penalty for non-compliance or audit failure.

In the simulation, administrative cost is proxied by the time and disagreement involved in committee-style deliberations:

$$\text{AdminCost}(s) = \theta_1 \cdot \text{DeliberationTime}(s) + \theta_2 \cdot \text{Disagreement}(s) \tag{4}$$

where DeliberationTime(s) and Disagreement(s) are set to 0 if the firm ignores the policy ($s = \text{Ignore}$). Varying fines, audit probabilities, and admissible comparison methods illustrates which policy designs close the simulated demand-based wage gap through wage adjustment rather than through non-compliance or strategic reporting.

References

[1] Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., & Zrnic, T. (2023). Prediction-powered inference. *Science*, 382(6671), 669-674.

[2] Fessler, P., & Kasy, M. (2019). How to use economic theory to improve estimators: Shrinking toward theoretical restrictions. *Review of Economics and Statistics*, 101(4), 681-698.

[3] Gouvernement du Québec. (1996). *Pay Equity Act*, chapter 43.