babl

OFFICIAL DOCUMENT
V1.0: **11/12/2024**

PREPARED BY BABL AI INC.

SUMMARY OF BIAS AUDIT RESULTS

Audit of **Gem's AI Ranking**
for New York City's Local Law 144

Presented to
Gem

Table of Contents

Letter from the Lead Auditor

From: **Shea Brown**
Lead Auditor
BABL AI Inc.
sheabrown@bablai.com

To: **Gem**
1 Post Street, Suite 1800
San Francisco, CA, 94104

Re: **Audit Opinion on Gem's AI Ranking**

11/12/2024

We have independently audited the bias testing assertions and related documentary evidence of Gem (the "Company") as of 11/12/2024, presented to BABL AI in relation to Company's AI Ranking in accordance with the criteria and audit methodology set forth in this report. The goals of this audit are to:

1. Determine whether the bias testing methodologies, controls, and procedures performed by Company satisfy the audit criteria (see Findings)
2. Obtain reasonable assurance as to whether the statements made by the Company, including the summary of bias testing results presented in this report, are free from material misstatement, whether due to fraud or error.

Note that the criteria presented in this report were constructed specifically to address the requirements of a "bias audit" outlined in NYC Local Law No. 144 of 2021. The model was audited as though it were an automated employment decision tool (AEDT) under NYC Local Law No. 144 of 2021, but we do not make any determination whether the model is, in fact, an AEDT under this law.

Company Responsibilities

It is the responsibility of Company representatives to ensure that bias testing and related procedures comply with the criteria outlined in this report. The Company representatives are responsible for ensuring that the documents submitted are fairly presented and free of misrepresentations, providing all resources and personnel needed to ensure an effective and efficient audit process, and providing access to evidential material as requested by the auditors.

BABL AI Responsibilities

It is the responsibility of the lead auditor to express an opinion on the Company's assertions related to the bias testing of the model. In light of the current absence of generally accepted standards for the auditing of algorithms and autonomous systems, our examination was conducted in accordance with the standards and normative references outlined in this report.

Those standards require that we plan and perform audit procedures to obtain reasonable assurance about whether the assertions referred to above 1) satisfy the audit criteria and 2) are free of material misstatement, whether due to error or fraud. Within the scope of our engagement, we performed amongst others the following procedures:

- Inspection of submitted documents and external documentation
- Interviewing Company employees to gain an understanding of the process for determining the disparate impact and risk assessment results
- Observation of selected analytical procedures used in Company's bias testing
- Inspection of the select samples of the bias testing data
- Inquiry of personnel responsible for governance and oversight of the bias testing and risk assessment

We believe that the procedures performed provide a reasonable basis for our opinion.

Independence

Our role as an independent auditor conforms to ForHumanity and Sarbanes-Oxley definitions of Independence. Fees associated with this contract are for the provision of the service to assess compliance. The payment of fees is unrelated to the decision rendered. Our decision is grounded solely in the criteria presented below.

Opinion

In our opinion, based on the procedures performed and the evidence received to obtain assurance, the bias testing and results presented by Company, as of 11/12/2024, is prepared, in all material respects, in accordance with the criteria outlined below.

Sincerely,

Shea Brown

Shea Brown
Lead Auditor, BABL AI Inc.

System Description

BABL AI was engaged to audit Gem's AI Ranking (the "System"). The system is powered by a large language model and scores a candidate's resume based on a set of evaluation criteria defined by the user (e.g., recruiters). A final score is subsequently produced by the system and displayed to the user in a ranked list.

The final score ranges from 0–100% and is computed based on the weights of the criteria provided by the user. The median score for the dataset was used to compute the "scoring rate" for candidates of various self-declared demographic groups. The scoring rate for each demographic group is displayed in the summary of the Disparate Impact results in the Findings section.

Audit Summary

Background

New York City Local Law No. 144 of 2021 requires yearly "bias audits" for automated employment decision tools (AEDTs) used to substantially assist or replace decisions in hiring or promotion. Specifically, the law states that (1) the bias audit must "assess the [AEDTs'] disparate impact" on certain persons, (2) the audit must be conducted by an "independent auditor … no more than one year prior to the use", and (3) a "summary of the results of the most recent bias audit … [must be] made publicly available on the website of the employer or employment agency." The audit outlined in this document has been conducted to satisfy the law's requirement for a bias audit only, and does not include other requirements such as candidate notifications. This report does not make any determination whether the model under this audit is, in fact, an automated employment decision tool as defined under NYC Local Law 144, or not.

Auditor Responsibilities

It is the responsibility of BABL AI auditors to:

1. **Obtain reasonable assurance** as to whether the statements made by the auditee are free from material misstatement, whether due to fraud or error,
2. **Determine whether the statements** made by the auditee provide sufficient evidence that the audit criteria (see Findings) have been satisfied, and
3. **Issue an auditor's report** that includes an opinion.

As part of an audit in accordance with good auditing practice, BABL AI exercises professional judgment and maintains professional skepticism throughout the audit. Specifically, BABL AI auditors identify and assess the risks of material misstatement in documents provided by the auditee, perform audit procedures responsive to those risks, and obtain audit evidence that is sufficient and appropriate to provide a basis for our opinion, per Public Company Accounting Oversight Board (PCAOB)'s Auditing Standard 1105 on Audit Evidence,[1] where applicable. In addition, this audit report follows International Standard on Assurance Engagements (ISAE) 3000's guidelines on Assurance Report, where applicable.[2]

BABL AI is also responsible for maintaining auditors' independence and objectivity to ensure the integrity of the opinion and certification provided. BABL AI as an organization, and all employee and contract auditors, adhere to strict independence as codified by the

[1] AS 1105: Audit Evidence
[2] ISAE 3000: Assurance Engagements Other Than Audits or Reviews of Historical Financial Information

Sarbanes–Oxley Act of 2002[3] and the ForHumanity's Code of Ethics.[4] In addition, BABL AI Lead Auditors are ForHumanity Certified Auditors under NYC AEDT Bias Audit.[5] For more details about our methodology and process, see Appendix – Audit Methodology.

Scope & Objective

Audit Section	Audit Objective
Disparate Impact Quantification	To ensure that the auditee has conducted sufficient testing of their model to "assess the tool's disparate impact on persons of any component 1 category," – i.e., race and gender – as the minimal requirement for a bias audit under Local Law 144 of 2021.
Governance	To ensure that effective internal governance exists to own, manage, and monitor risks related to bias and fairness.
Risk Assessment	To ensure that risks of the model that potentially contribute to bias have been rigorously identified, acknowledged, and assessed.

Out of Scope

1. The audit did not ensure the sufficient testing of the tool's disparate impact on any other protected class beyond race/ethnicity and gender
2. The audit did not certify that the model is "bias-free"
3. The audit is not intended for compliance purposes for any legislation other than the NYC AEDT law

[3] Sarbanes–Oxley Act of 2002
[4] ForHumanity Certified Auditor Code of Ethics
[5] ForHumanity NYC Bias Audit

Conclusions

Our opinions for the bias audit of **AI Ranking** are as follows:

Audit Section	Opinion
Disparate impact quantification	**PASS**
Governance	**PASS**
Risk assessment	**PASS**
Overall	**PASS**

Findings

Note: *The information disclosed under each criterion is not documentary evidence.*

Disparate Impact Quantification

Audit Criteria	Opinion
Q.A. Components: The model to be tested for disparate impact shall be defined. Q.A.1. Where the model comprises more than one automated component, evidence shall show appropriate definition of the model.	**PASS**

Components or combinations of components that were tested:N/A

Q.B. Testing dataset: The dataset on which disparate impact was quantified shall be defined and characterized. Q.B.1. Evidence shall show justification for why the selected dataset was appropriate for disparate impact testing. Q.B.2. Where test data as defined in § 5-300 was used, evidence shall show a. justification for not using historical data, b. that historical data is not sufficient to perform a statistically significant disparate impact testing, and c. the methodology by which test data was collected Q.B.3. Where disparate impact testing was not completed by BABL, evidence shall show a. that the most recent testing was conducted less than one year prior to the start date of this audit, or after a major update to the model, unless the update was more than one year prior to the start date of this audit, in which case, evidence shall show b. justification for why such testing was still appropriate. Q.B.4. Evidence shall show that the data used in the testing was within one year of the start date of the disparate impact testing.	**PASS**

Testing conducted by: Gem
Date of last testing: Sep 2024
Time span of data: Dec 2018 – Sep 2024

Q.C. Disparate-impact quantifiable PCVs: PCVs that can be quantified using the testing dataset shall be defined. Q.C.1. Evidence shall identify PCVs that were quantifiable in regard to disparate impact. Q.C.2. Evidence shall show that the PCVs that can be quantified include at the least: race, and gender. Q.C.3. Evidence shall disclose the method by which PCV data was collected. Q.C.4. Evidence shall identify and disclose PCVs that were not quantified in regard to disparate impact. Q.C.5. Where PCV data was inferred, evidence shall a. identify the method by which PCV data was inferred, and b. show justification for why the selected method of PCV inference was appropriate.	**PASS**

PCVs for which disparate impact was quantified:

1. Gender
2. Race/ethnicity

PCVs for which disparate impact was not quantified:

1. Age
2. Immigration or citizenship status
3. Disability status
4. Marital status and partnership status
5. National origin
6. Pregnancy and lactation accommodations
7. Religion/creed
8. Sexual orientation
9. Veteran or Active Military Service Member status

Q.D. Positive vs. negative outcome: Where the selection rate method was used, positive and negative outcomes of the model shall be clearly defined. Q.D.1. Evidence shall show justification for why the selected definition of positive outcome was appropriate. Q.D.2. Where thresholding is used, evidence shall show justification for why the level/levels of threshold to determine positive vs. negative outcomes was/were appropriate. Q.D.3. Evidence shall identify and disclose a. all user-configurable settings, b. whether each setting affects positive outcomes, and for all settings that affect outcomes, c. their extents of user configurability, d. their default values, and e. justification for why such default values were appropriate. Q.D.4. Evidence shall disclose the user-configurable settings and combinations of settings on which disparate impact was tested.	**PASS**

Positive outcome: N/A, due to the use of scoring rate method

User-configurable settings that can affect scoring rate:

1. Criteria for resume scoring
2. Weights for criteria components

Settings on which disparate impact was tested: The default weights for the criteria were used for testing.

Q.E. Selection rate or scoring rate: A metric corresponding to selection rate or scoring rate shall be defined. Q.E.1. Where the selection rate method was used, evidence shall show that the selection rate of a group was defined as the ratio of positive outcome to all outcomes for that group. Q.E.2. Where the scoring rate method was used, evidence shall show that the scoring rate of a group was defined as the rate at which that group receives a score from the AEDT above the median score of the sample	**PASS**

Method of quantifying disparate impact: Scoring rate, as defined by the proportion of a demographic group having a score above the median score of the population.

Q.F. Favored, disfavored groups: Favored and disfavored groups shall be defined, for all PCVs. Q.F.1. Evidence shall show that favored and disfavored groups were defined according to selection rates or scoring rates ordered by PCV. Q.F.2. Evidence shall show that the groups pertaining to race and ethnicity satisfy § 60-3.4 B in the EEO guidelines. Q.F.3. Where the groups pertaining to race and ethnicity do not satisfy EEO guidelines, evidence shall show justification for why EEO grouping was not used, and the appropriateness of any substituted groupings. Q.F.4. Evidence shall show that the groups pertaining to gender contain at least "Male" and "Female". Q.F.5. Evidence shall show intersectional groups containing all permutations of gender and race/ethnicity group combinations. Q.F.6. Where race/ethnicities and genders are not known for a sample of candidates assessed by the AEDT, evidence shall disclose its sample size.	**PASS**

Q.G. Impact ratio: Impact ratios shall be disclosed for all disfavored groups, for all PCVs. Q.G.1. Where an impact ratio for a disfavored group is below 0.8, evidence shall show justification for why the disfavored group is disadvantaged. Q.G.2. Evidence shall show results of uncertainty analysis (e.g., standard error for the mean) or error propagation of impact ratios in the form of errors or error bars. Q.G.3. Where PCV data was inferred, evidence shall show that systematic errors due to PCV inference were properly propagated in impact ratio calculations. Q.G.4. Where a gender, race/ethnicity, or intersectional group was excluded from impact ratio calculation due to its size being below 2% of the total sample size of each analysis, evidence shall show a. justification for the exclusion of such group b. the sample size of such group, and c. the selection rate or scoring rate of such group	**PASS**

Non-intersectional, Gender, sorted by Scoring rate

	N applicants	Scoring rate	Impact ratio
Male	7,903	0.502	1.000
Female	3,946	0.465	0.927

Non-intersectional, Race/ethnicity, sorted by Scoring rate

	N applicants	Scoring rate	Impact ratio[6]
Native Hawaiian or Pacific Islander	21	0.571	N/A
Two or more races	322	0.500	1.000
Asian	8,438	0.496	0.993
White	1,649	0.481	0.962

[6] N/A refers to the demographic group representing less than 2% of the total N applications in the table.

	N applicants	Scoring rate	Impact ratio[6]
Hispanic or Latino	512	0.451	0.902
Black or African American	420	0.433	0.867
Native American or Alaskan Native	19	0.421	N/A

Intersectionals

			N applicants	Scoring rate	Impact ratio[7]
Hispanic or Latino	Male		368	0.451	0.886
	Female		137	0.445	N/A
Non-Hispanic or Latino	Male	White	1,108	0.500	0.982
		Asian	5,558	0.509	1.000
		Black or African American	292	0.425	0.834
		Native American or Alaskan Native	15	0.400	N/A
		Native Hawaiian or Pacific Islander	16	0.563	N/A
		Two or more races	216	0.481	0.946
	Female	Asian	2,843	0.472	0.927
		White	536	0.444	0.872
		Black or African American	124	0.452	N/A
		Native American or Alaskan Native	4	0.500	N/A

[7] N/A refers to the demographic group representing less than 2% of the total N applications in the table.

			N applicants	Scoring rate	Impact ratio[7]
		Native Hawaiian or Pacific Islander	5	0.600	N/A
		Two or more races	97	0.536	N/A

Note: Data on these applicants was not included in the calculations above:

1. 4,097 applicants with an unknown gender category, and
2. 4,565 applicants with an unknown race/ethnicity category

Q.H. Statistical significance: Where the selection rate method was used, statistical significance calculation shall satisfy UGESP guidelines. Q.H.1. Evidence shall show that statistical significance was calculated using the Two Independent-Sample Binomial Z-Test for sample sizes of 30 or more, and using the Fisher's Exact Test for sample sizes of fewer than 30.	**N/A**[8]

[8] Due to the use of scoring rate method

Governance

Audit Criteria	Opinion
G.A. Accountable party for disparate impact risks: The auditee shall have a party who is accountable for risks related to disparate impact. G.A.1. Evidence should show that the accountable party is a committee, but may also show that the accountable party is a single individual. G.A.2. Evidence shall clearly show that risks related to disparate impact are owned and managed by the accountable party.	**PASS**

Accountable party: Gem AI Governance Group
Contact information: Matt Flairty, mflairty@gem.com
Role in the auditee organization: Legal and Compliance

G.B. Defined duties of the accountable party: Duties of the party accountable for disparate impact risks shall be clearly defined. G.B.1. Evidence shall show that such duties pertain to the ownership, management, and monitoring of disparate impact risks. G.B.2. Evidence shall show that the accountable party has influence over product changes per effective challenge in Federal Guidance on Model Risk Management.	**PASS**

G.C. Documentation pertaining to duties carried out: The auditee shall provide evidence that the defined duties of the party accountable for disparate impact risks are carried out. G.C.1. Evidence shall show that the defined duties were carried out prior to the start date of this audit.	**PASS**

Risk Assessment

Audit Criteria	Opinion
R.A. Completion: The auditee shall have completed a risk assessment of the model. R.A.1. Evidence shall show that a risk assessment or an equivalent analysis was completed less than one year prior to the issuance date of this audit.	**PASS**

Evidence of Risk Assessment completion: Risk assessment, governance group meeting minutes and verbal testimony from the accountable party.

R.B. Identification of risks: Risk assessment shall show identification of relevant risks related to bias. R.B.1. Evidence shall show identification of risks related to various biases along all stages of the AI life cycle as listed in NIST Standard for Identifying and Managing Bias in Artificial Intelligence. R.B.2. Evidence shall show awareness of the parties potentially affected by the decisions made along all stages of the AI life cycle.	**PASS**

R.C. Evaluation of risks: Risk assessment shall demonstrate appropriate evaluation of relevant risks. R.C.1. Evidence shall show that the identified risks are assessed from the perspectives of multiple affected external and internal stakeholders, with justifications for the extent of and mechanism by which such risks affect these stakeholders. R.C.2. Evidence shall show that the identified risks are assessed in a sufficiently rigorous manner, using a quantitative and/or qualitative evaluation scheme, and along multiple dimensions, such as but not limited to likelihood of harm and severity of harm. R.C.3. Evidence shall show justification for the provided evaluation of risks.	**PASS**

Appendix

Audit Methodology

The Criterion Audit

The BABL AI audit framework is the *Criterion Audit Framework*,[9] defined as "a criteria-based independent external evaluation of an algorithmic system conducted by an auditor to determine whether the given system meets the requirements set by a normative framework." A criterion audit is modeled after the financial auditing practice, and is distinguished from other commonly used forms of assessment of algorithms, such as internal audits, critical third-party audits, and risk or impact assessments. The audit framework contains three main phases:

1. **Scoping** – The auditor conducts a preliminary survey of the auditee's algorithm to gain a full understanding to contextualize documentary evidence
2. **Evaluation & Verification** – The auditee submits documentation containing evidence demonstrating satisfaction of the audit criteria which the auditors evaluate and verify.
3. **Certification** – If the auditee is determined to pass the audit criteria, the auditor drafts the auditor's report and certifies the auditee's algorithm.

Evaluation & Verification

The procedure for all BABL AI auditors to conduct a criterion audit follows the guidelines set forth in the Public Company Accounting Oversight Board (PCAOB)'s Auditing Standard 1105 on Audit Evidence, where applicable. Specifically, the auditors:

1. **Obtain audit claims and statements** from the auditee's submitted documentation which either support or contradict the criteria and sub-criteria,
2. **Evaluate the claims and statements** in regard to satisfying the criteria and sub-criteria, based on the *sufficiency* and *appropriateness* of the evidence, and
3. **Verify that the claims and statements** made by the auditee are free from material misstatement, whether due to fraud or error.[10]

[9] Lam, K., Lange, B., Blili-Hamelin, B., Davidovic, J., Brown, S. & Hasan, A. (2024). A Framework for Assurance Audits of Algorithmic Systems. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24. ACM, June 2024. doi: 10.1145/3442188.3445924.

[10] "Reasonable assurance" is a high level of assurance but is not a guarantee that an audit conducted in accordance with good auditing practice always detects a material misstatement when it exists. Misstatements can arise from fraud or error and are considered material if, individually or in aggregate, they could reasonably be expected to influence the decisions of stakeholders taken based on these statements.

In addition, evaluation and verification of claims and statements may involve requesting additional supporting documentary evidence, and/or interviewing those responsible for the governance of the algorithm, other relevant employees of the auditee organization, or other third parties referenced in the submitted documentation.

At the end, the auditors reach an audit opinion based on:

1. The sufficiency and appropriateness of the audit evidence, and
2. The risk of material misstatement of the audit evidence.

Terminologies & Definitions

Term	Abbrev	Definition
automated employment decision tool	AEDT	“any computational process, derived from machine learning, statistical modeling, data analytics, or artificial intelligence, that issues simplified output, including a score, classification, or recommendation, that is used to substantially assist or replace discretionary decision making for making employment decisions that impact natural persons.” – see § 20-870 of the Code and § 5-300 of the adopted rule for full definition
disfavored group		any gender or race/ethnicity group not having the highest selection rate or average score
disparate impact or adverse impact		“a selection rate for any race, sex, or ethnic group which is less than four-fifths (⅘) (or 80%) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact” – see § 60-3.4.D of UGESP (1978) for full definition
error propagation		calculation or computation of a variable's uncertainty that is dependent on another variable's uncertainty
favored group		the gender or race/ethnicity group having the higher selection rate or average score compared to the other groups
impact ratio		“either (1) the selection rate for a category divided by the selection rate of the most selected category or (2) the scoring rate for a category divided by the

Term	Abbrev	Definition
		scoring rate for the highest scoring category. " – see § 5-300 of the adopted rule for full definition
scoring rate		"the rate at which individuals in a category receive a score above the sample's median score, where the score has been calculated by an AEDT"
justification		a compelling reason that illuminates the issue and carries normative force, as opposed to solely explanatory power
positive outcome		the basis for selection rate, the favorable outcome for a candidate from the use of the model, such as being selected to move forward in the hiring process or assigned a classification by an model
protected category variables	PCV	defined per jurisdiction, equivalent to protected class, including but not limited to: race/ethnicity, age, gender, religion, ability or disability, sexual orientation, color, nation of origin, socioeconomic class
risk assessment		an assessment of the risk that the use of the algorithm negatively impacts the rights and interests of stakeholders, with a corresponding identification of situations of the context and/or features of the algorithm which give rise or contribute to these negative impacts[11]
selection rate		"the rate at which individuals in a category are either selected to move forward in the hiring process or assigned a classification by an AEDT" – see § 5-300 of the adopted rule for full definition
testing dataset		the dataset used to test for or quantify disparate impact
uncertainty analysis		calculation or computation to quantify the uncertainty of a variable, outputting errors or error bars

[11] Hasan, A., Brown, S., Davidovic, J., Lange, B., & Regan, M. (2022). Algorithmic Bias and Risk Assessments: Lessons from Practice. *Digital Society*, 1(1). doi: 10.1007/s44206-022-00017-z.