Protocole d'Audit des Biais Algorithmiques en Recrutement

Mesure de la Robustesse Statistique des Extractions et Évaluations par LLM

Abdoul-Aziz TOURE
SEMANTIKMATCH
abdoul@semantikmatch.com

5 janvier 2026

Résumé

L'automatisation du recrutement par LLM (Large Language Models) implique une chaîne complexe : lecture de documents hétérogènes (CV, Bulletins) et extraction structurée. Ce protocole définit une méthodologie expérimentale rigoureuse pour isoler les biais d'extraction liés à la forme (OCR, mise en page) et aux attributs démographiques (genre, origine). L'approche inclut un plan de validation statistique complet (Tests Chi-2, Mann-Whitney, ANOVA) pour garantir que les écarts de performance observés sont significatifs et non aléatoires.

Table des matières

1 Objectifs de l'Étude

L'objectif est de mesurer la **Fidélité d'Extraction** des modèles sur des documents (CV et Bulletins) présentant des variations contrôlées.

Nous cherchons à répondre aux questions suivantes :

1. **Biais Démographique d'Extraction :** Le modèle fait-il plus d'erreurs de lecture (omissions, hallucinations) quand le document contient un prénom féminin ou à consonance étrangère ?
2. **Biais de Format/Origine :** Le modèle extrait-il aussi fidèlement les notes d'un bulletin issu d'un système éducatif non-occidental (ex : note/20) que celles d'un système standard (GPA/4.0) ?

2 Méthodologie Expérimentale

2.1 Approche par "Vérité Terrain" (Ground Truth)

Nous utilisons une approche supervisée :

1. Création d'un **Profil Logique** (JSON Gold) contenant les données parfaites.
2. Génération de documents (CV et Bulletins) représentant ce profil, en injectant des variations (Identité, Mise en page).
3. Extraction par le LLM.
4. Comparaison mathématique entre le JSON Gold et le JSON Extrait.

2.2 Variables Testées

— **Variable A : Identité (Démographie)** : Genre (H/F), Origine (Proxy Prénom/Lieu), Âge (Dates).

— **Variable B : Typologie Documentaire** : Support (CV/Bulletin), Qualité (Natif/Scan OCR).

3 Plan d'Analyse et Validation Statistique

Afin d'assurer la robustesse des résultats, les analyses descriptives (moyennes) seront complétées par des tests d'inférence statistique. Le seuil de significativité est fixé à $\alpha = 0.05$.

3.1 Synthèse des Tests Statistiques

Table 1 – Matrice des tests statistiques par typologie de données

Hypothèse à tester	Type de Données	Test Statistique	Condition d'application
Biais d'Extraction (Taux d'erreur, Omissions)	Catégorielles (Succès / Échec)	**Chi-carré** (χ^2) Test Exact de Fisher	Effectifs théoriques > 5 Effectifs faibles
Impact du Format (Score de Précision OCR)	Quantitatives (Score 0-100%)	**Test de Mann-Whitney U** Test t de Student	Distribution non-normale Distribution normale
Biais Démographique (Score Évaluation 1-10)	Ordinales/Quantitatives (Note attribuée)	**Kruskal-Wallis** ANOVA à 1 facteur	Plus de 2 groupes Homogénéité des variances
Interactions (Genre × Format)	Quantitatives (Score)	**ANOVA Factorielle** (Two-way)	Vérifier si les biais se cumulent (ex : Femme + Scan flou)
Impact Métier (Lien Erreur → Note)	Quantitatives (Deltas)	**Corrélation de Spearman** Corrélation de Pearson	Relation monotone Relation linéaire stricte

3.2 Détail des Procédures Statistiques

3.2.1 Analyse des Fréquences d'Erreur (χ^2)

Pour vérifier si l'origine du candidat influence le taux de succès de l'extraction.

— H_0 : Le taux d'erreur est indépendant de l'origine du candidat.

— H_1 : Il existe une dépendance significative entre l'origine et le taux d'erreur.

3.2.2 Analyse des Scores d'Extraction (Mann-Whitney U)

Utilisé pour comparer la précision d'extraction entre deux formats (ex : PDF Natif vs Scan). Nous privilégions le test non-paramétrique de Mann-Whitney car les scores de précision (souvent proches de 100%) ne suivent généralement pas une distribution normale (effet plafond).

3.2.3 Analyse des Biais de Notation (ANOVA / Kruskal-Wallis)

Pour vérifier si le score moyen attribué (sur une échelle 1-10) varie selon le groupe démographique. Si l'ANOVA révèle une différence significative ($p < 0.05$), des tests **post-hoc de Tukey** seront réalisés pour identifier précisément quels groupes diffèrent (ex : Junior vs Senior).

3.2.4 Correction de Bonferroni

Puisque de multiples hypothèses sont testées simultanément sur le même jeu de données (comparaisons multiples), nous appliquerons une correction de Bonferroni pour ajuster le seuil α et éviter les "Faux Positifs" statistiques :

$$\alpha_{corrigé} = \frac{0.05}{\text{Nombre de tests réalisés}}$$

4 Métriques de Mesure (KPIs)

Nous ne mesurons pas le "score" du candidat, mais la **qualité du parsing**.

4.1 1. Taux d'Exactitude (Accuracy)

Pour les champs factuels (Nom de l'école, Intitulé du diplôme, Entreprise) :

$$Accuracy = \frac{\text{Nb de champs extraits strictement identiques au Gold}}{\text{Nb total de champs}}$$

4.2 2. Précision Numérique (Sur les Bulletins)

Pour les notes et moyennes :

$$\Delta_{Note} = |\text{Note}_{Extraite} - \text{Note}_{Reelle}|$$

4.3 3. Stabilité (Coefficient de Variation)

Mesure de la robustesse déterministe du modèle.

$$CV = \frac{\sigma}{\mu} \times 100$$

Si $CV > 10\%$, l'extracteur est considéré comme instable.

5 Schéma JSON de Sortie

Listing 1 – Format cible pour l'extraction

```
{
  "identity_extraction": {
    "detected_name": "String",
    "detected_email": "String",
    // Le modele a-t-il correctement capte l'adresse ?
    "address_components": {
        "city": "String",
        "country": "String"
    }
  },
  "education_extraction": [
    {
      "institution_name": "String",
      "degree_title": "String",
      "graduation_date": "YYYY-MM",
      // Extraction specifique Bulletin
      "transcript_data": {
          "overall_grade": "Number/String",
          "grade_scale_detected": "String (ex: 0-20, A-F)",
          "courses": [
              {"name": "Maths", "grade": "14"}
          ]
      }
    }
  ]
}
```

6 Conclusion

L'intégration de ces tests statistiques permet de dépasser la simple observation de tendances. Elle fournit un cadre de preuve mathématique nécessaire pour valider ou rejeter l'hypothèse selon laquelle les LLM introduisent des biais structurels ou démographiques lors du traitement automatisé des dossiers candidats.