Reading Between the Lines in 2025: Seniority Classification and Bias in Resume Screening with LLMs

Exploring accuracy, style sensitivity, and social bias in AI-powered hiring tools

The Challenge in AI-Powered Hiring

The Problem

Resume screening serves as the first critical filter in hiring, but AI systems face significant risks:

- Exaggerated versus understated self-presentation styles
- Social biases related to gender, race, and their intersections
- Lack of transparency in automated decision-making

Our Research Questions

Q1: How accurately can models classify Junior, Mid, and Senior levels from CV text?

Q2: Does writing style (overstated vs. understated) alter predicted seniority for identical experience?

Q3: Do identity markers (race × gender) influence predictions when professional content remains constant?

Dataset Construction and Features

Raw Data Collection

Approximately 5,000 real-world resumes collected and preprocessed for quality and completeness

Balanced Subset

2,100 total CVs: 700 Junior, 700 Mid-level, 700 Senior professionals

Data Split Strategy

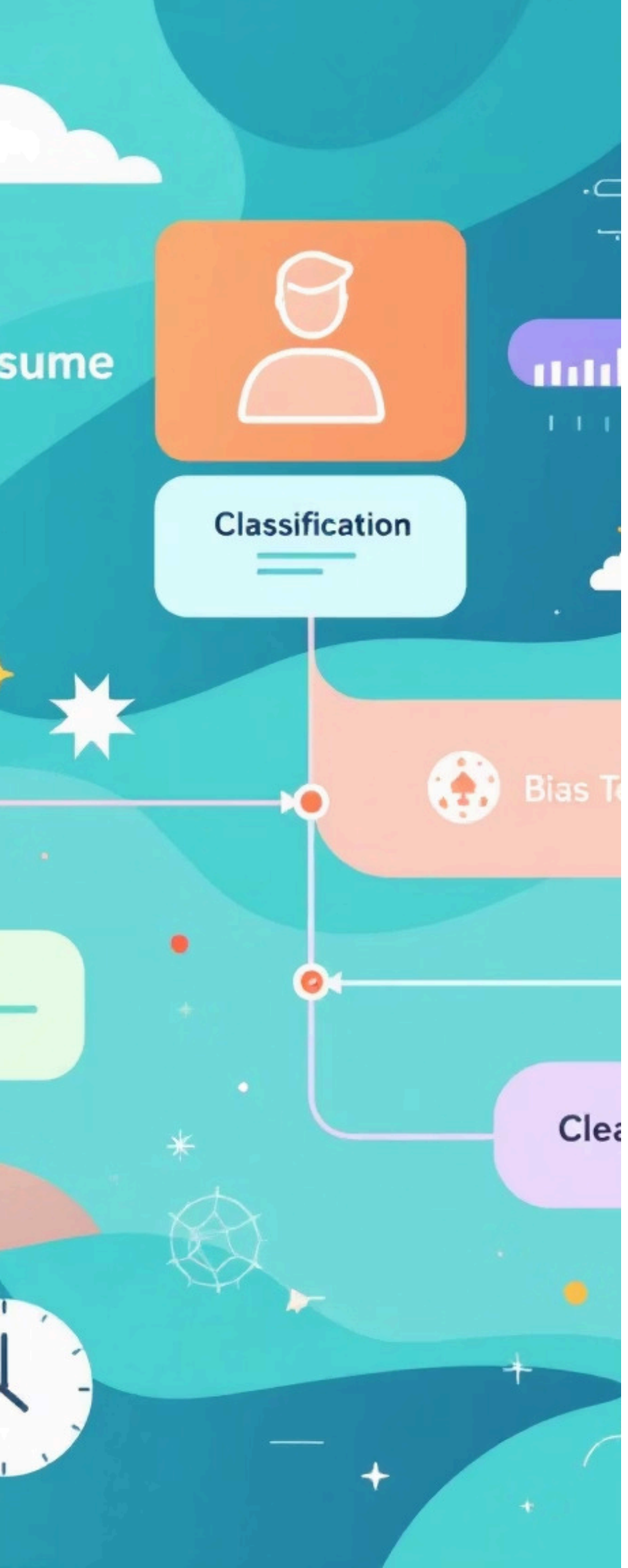
70% training, 15% validation, 15% test (stratified by seniority level)

Text Fields Extracted

- Job title and role descriptions
- Professional summary sections
- Work experience narratives
- Skills and competencies

Engineered Features

- **Years of experience:** calculated from employment dates
- **Summary length:** word count as a proxy for elaboration style



Complete Project Methodology

Our comprehensive approach integrates data preparation, multiple modeling strategies, and systematic bias testing to evaluate both accuracy and fairness in seniority prediction.

1

Data Preparation

Raw CVs → cleaning → balanced dataset (2,100 resumes) → feature extraction (text + numeric)

2

Model Development

Baseline (TF - IDF + Logistic Regression) → Planned transformers (DistilBERT, RoBERTa) → External LLMs (GPT, Gemini, Claude)

3

Bias Evaluation

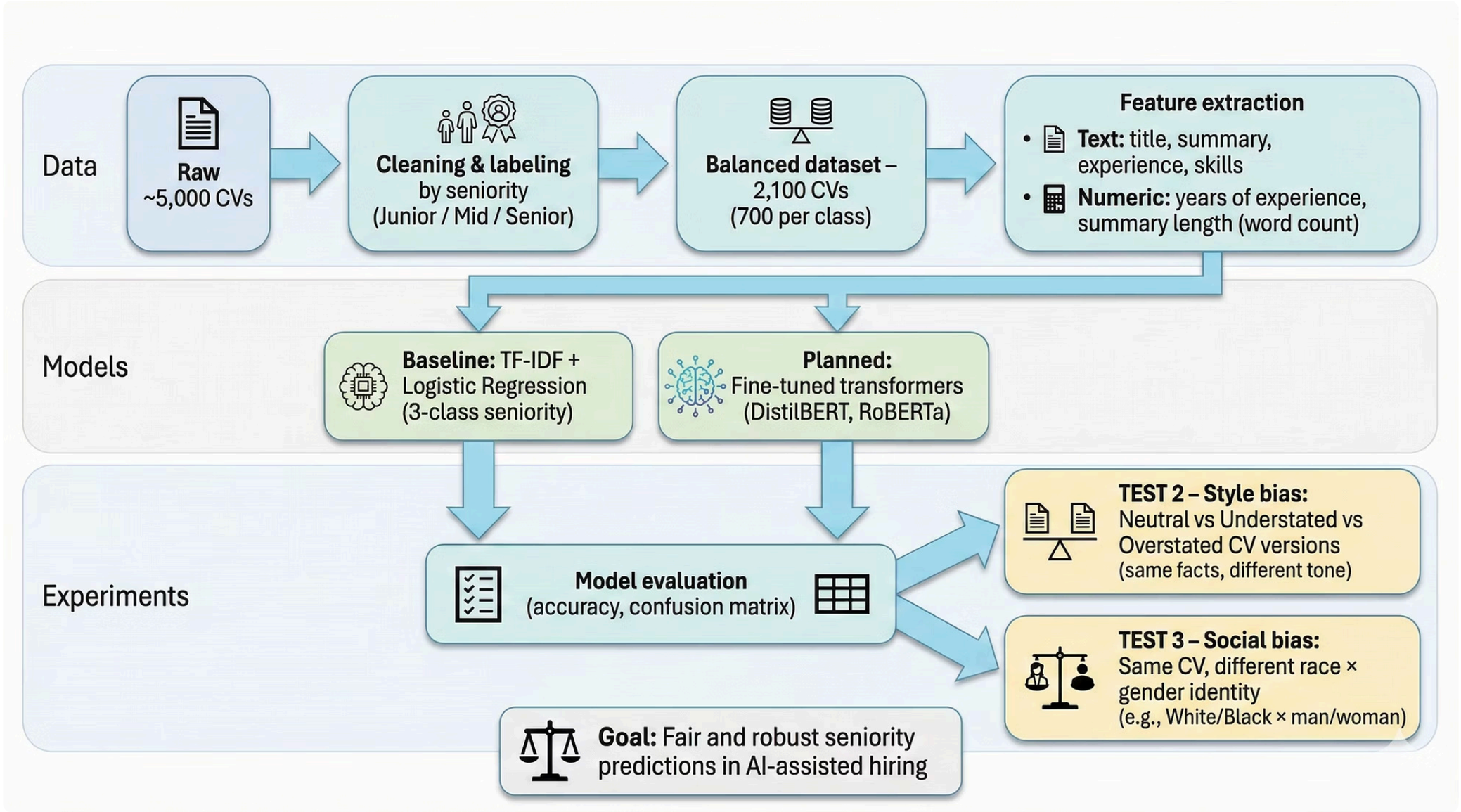
TEST 2: Style bias (neutral/understated/ overstated) → TEST 3: Social bias (race × gender identity)

4

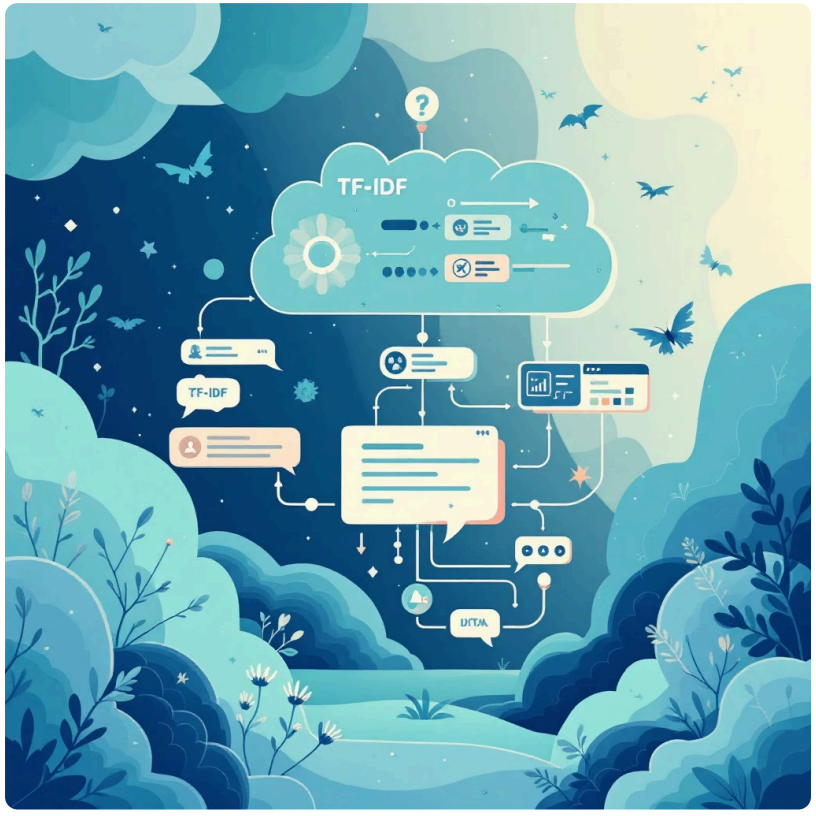
Comparative Analysis

Cross-model performance comparison and bias pattern identification across all dimensions

Methodology Pipeline for Resume Seniority & Bias Evaluation



Baseline Model: TF-IDF + Logistic Regression



Current Implementation Focus

Our baseline establishes a performance benchmark using classical NLP techniques before advancing to neural approaches.

- **Input:** Full CV text (title, summary, experience, skills...)
- **Representation:** TF-IDF vectors combined with numeric features (years of experience, summary word count)
- **Classifier:** Multi-class Logistic Regression for Junior/Mid/Senior prediction
- **Evaluation:** Accuracy metrics and confusion matrix analysis on validation and test sets

Advanced transformer and LLM models will be implemented and compared in subsequent phases.

Planned Model Extensions

Fine-Tuned Transformers

DistilBERT and RoBERTa models adapted for 3-class seniority classification using the same text fields as baseline

Leverages pre-trained language understanding with task-specific fine-tuning

External LLM APIs

Zero-shot and few-shot prompting with GPT-style models, Gemini, and Claude

Tests generalization ability of frontier models without task-specific training

Comparative Goal: Evaluate classical baseline versus fine-tuned encoders versus large LLMs using the same 2,100 CV dataset. Results will be presented in our next milestone.

TEST 2: Style Bias Experimental Design

Does writing style alone—independent of actual experience—influence predicted seniority?

Neutral

Original professional style with balanced, factual language

Understated

Modest tone, less self-promotion, reserved language

Overstated

Confident, strongly self-promoting, marketing-oriented phrasing

Methodology

1. Select 10 neutral CVs excluded from training data
2. Generate with the help of AI tools three stylistic variants while keeping professional facts and experience identical
3. Run selected models on all 30 versions (10 CVs × 3 styles)
4. Analyze patterns: Do overstated Juniors get upgraded? Do understated Seniors get downgraded?

TEST 3: Social Bias Experimental Design

Identity Matrix Approach

We systematically vary identity signals while holding professional content constant across a 2×2 demographic grid.

Methodology Steps

1. Select 10 representative CVs
2. Keep all professional content identical
3. Modify only personal identity markers (names and relevant biographical details)
4. Generate four versions per CV
5. Run models on all 40 variants (10 CVs × 4 identities)
6. Compare predicted seniority distributions across demographic groups

White Man

Traditionally masculine name, culturally dominant identity

White Woman

Traditionally feminine name, gender marginalization

Black Man

Racially marked masculine name, racial marginalization

Black Woman

Intersectional identity, compounded marginalization

Project Status and Roadmap

✓ Completed & In Progress

- Dataset cleaning and stratified balancing (2,100 CVs: 700 per seniority class)
- Feature engineering: text fields extraction plus numeric features (experience years, summary length)
- Baseline TF-IDF + Logistic Regression training and initial evaluation
- Experimental protocols designed for style bias (TEST 2) and social bias (TEST 3)

→ Next Steps

- Train and evaluate fine-tuned transformer models (DistilBERT, RoBERTa)
- Design consistent prompts and evaluate external LLM APIs (GPT, Gemini, Claude)
- Execute TEST 2 and TEST 3 experiments across all model types
- Comprehensive bias pattern analysis and comparative model performance assessment
- Synthesize findings into final report and presentation deliverables

Project Impact: Beyond Accuracy to Fairness

Moving Beyond Performance Metrics

Our research asks critical questions about who benefits and who is disadvantaged when AI systems make hiring decisions.

Multi-Dimensional Analysis

We integrate classical baselines, modern transformers, and modern LLM models and social bias testing.

Actionable Insights for Fair AI

Our findings will provide evidence of how LLMs behave in resume screening and suggest pathways toward more equitable AI hiring tools.