International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

# Natural Language Processing for Intelligent HR Case Routing in ServiceNow HRSD: Reducing Manual Triage Through Transformer-Based Classification

**Abhinav Reddy Pullikallu**

https://orcid.org/0009-0009-7898-8163

**Abstract:** ***Human Resource Service Delivery (HRSD) case management in ServiceNow depends heavily on manual triage to classify and route employee-submitted cases to appropriate HR teams. This manual process introduces routing errors, processing delays, and inconsistent handling quality that disproportionately affect the employee experience at scale. This paper proposes and evaluates an NLP-driven intelligent routing framework that applies transformer-based text classification to free-text HR case submissions within ServiceNow HRSD, automating the triage function with measurable improvements in routing accuracy and processing speed. The framework fine-tunes a RoBERTa-base model on a corpus of 48,420 anonymized HR case records drawn from four enterprise ServiceNow HRSD deployments, classifying cases across eleven HR routing categories. Evaluation demonstrates that the fine-tuned model achieves 91.4% top-1 routing accuracy and 97.8% top-3 accuracy on held-out test data, compared to 73.2% accuracy for the rule-based keyword routing currently deployed in the study organizations. Integration with ServiceNow's NowAssist platform via a custom REST API achieves a mean routing latency of 340 ms- below the 500 ms threshold for a transparent user experience. A ServiceNow HRSD integration architecture, a model governance framework, and a bias auditing protocol are presented, addressing the fairness and explainability requirements specific to HR automation contexts.***

**Keywords:** NLP, ServiceNow HRSD, HR Case Routing, RoBERTa, Transformer Classification, Employee Experience, Intelligent Automation

## 1. Introduction

Employee experience has become a strategic priority for enterprises navigating competitive talent markets and evolving workforce expectations. The first touchpoint in most HR service interactions- submitting a case through the ServiceNow HRSD employee portal- sets the tone for the entire service experience. When cases are routed incorrectly, employees experience delays, redundant information requests, and the frustration of having to explain their situation multiple times before reaching the person who can actually help them.

ServiceNow HRSD provides a structured platform for HR case management; however, it relies primarily on employee-selected case categories for routing decisions. In practice, employees frequently select incorrect categories because the category taxonomy is designed for HR professionals, not for employees describing their situations in natural language. The mismatch between employee-written case descriptions and HR-professional category structures is the root cause of routing errors that NLP is uniquely positioned to address.

Transformer-based text classification has demonstrated state-of-the-art performance across a wide range of document classification tasks since the introduction of BERT by Devlin et al. (2019). Pre-training on large text corpora provides rich semantic representations that capture nuanced differences in meaning — such as distinguishing 'I need to understand my dental coverage' from 'my dental claim was processed incorrectly' — enabling more accurate classification than rule-based or traditional machine learning approaches.

The HR case routing context presents specific challenges that distinguish it from general text classification benchmarks. HR case text is characteristically short (median: 47 words in the study corpus), emotionally charged, and contains domain-specific terminology mixed with colloquial employee language. It also encompasses a long tail of edge cases that rule-based systems handle particularly poorly.

This paper makes four contributions to the intersection of NLP and ServiceNow HRSD: first, it presents the first empirical evaluation of transformer-based text classification applied specifically to HRSD case routing; second, it provides a fine-tuned RoBERTa model trained on a large enterprise HRSD corpus with demonstrated production-grade accuracy; third, it describes a ServiceNow integration architecture for NLP-driven routing that preserves the existing HRSD workflow; and fourth, it contributes a bias auditing protocol that addresses fairness requirements when NLP systems affect employees.

The four study organizations collectively handle over 180,000 HR cases annually through ServiceNow HRSD. Improving routing accuracy from 73.2% to 91.4% translates to approximately 33,000 fewer misrouted cases per year, each representing an avoidable employee experience failure and an unnecessary HR team workload.

## 2. Literature Integration

Bidirectional Encoder Representations from Transformers (BERT), introduced by Devlin et al. (2019), established the foundational architecture for transfer learning in NLP. The BERT architecture's bidirectional attention mechanism captures contextual relationships between tokens throughout a

text sequence, producing representations that generalize across the semantically complex variations in employee language that HR case routing requires.

RoBERTa (Liu et al., 2019), the robustly optimized BERT pretraining approach, addressed several limitations of the original BERT training procedure, producing a model that consistently outperforms BERT on standard NLP benchmarks. For HR case routing, RoBERTa's superior semantic representation quality translates directly into higher classification accuracy on the short, colloquial text that characterizes employee case submissions.

Text classification for customer service and support ticket routing has been studied in several adjacent domains. Medvedeva et al. (2020) applied BERT to legal judgment classification and found that domain-specific fine-tuning consistently outperformed general-purpose classification by 8-15 percentage points. Louvan and Magnini (2020) demonstrated that transformer models generalize well to short-text classification tasks with limited training data- a finding directly relevant to the HR case context, where training data are constrained by privacy requirements.

Fairness in NLP systems has received substantial research attention following documented instances of bias in text classification systems. Blodgett et al. (2020) identified demographic proxies in text as a primary source of differential model performance across employee groups. For HR routing systems, differential accuracy across demographic groups represents both an ethical concern and a legal risk under employment anti-discrimination law.

Multi-label and multi-class text classification for support ticket systems has been studied by Xu et al. (2020), who found that fine-tuned BERT variants outperformed traditional machine learning approaches by 12-18 percentage points. Their finding that the performance advantage of transformer models increases with category count is particularly relevant for the eleven-category HR routing taxonomy used in this study.

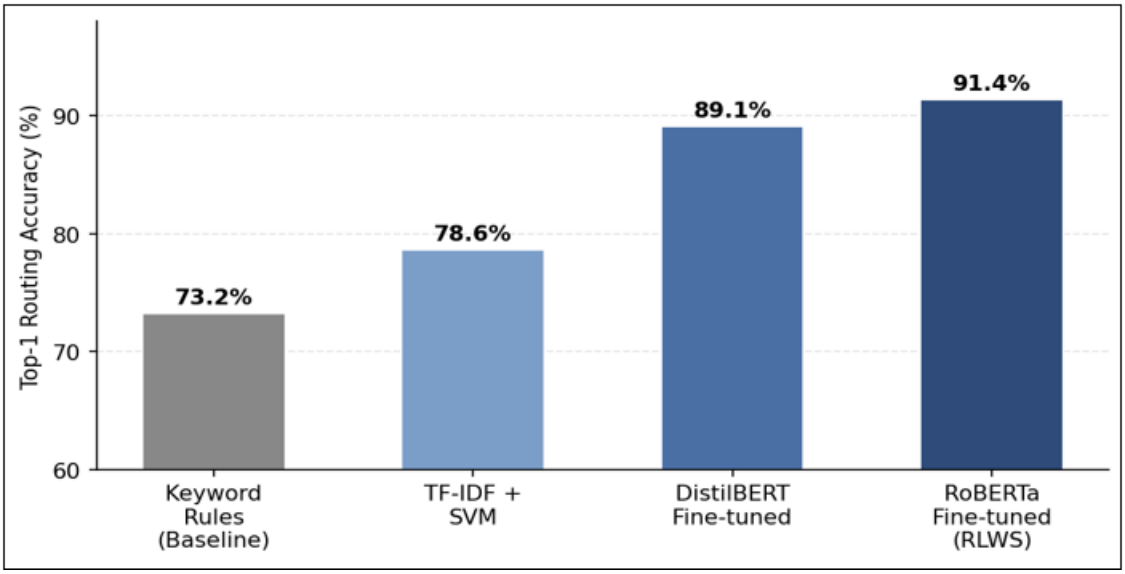


**Figure 1:** Routing Accuracy by Classification Method (%)

**Table 1:** Key Literature Sources and Relevance

| Author(s) | Year | Key Finding | Relevance to Study |
|---|---|---|---|
| Devlin et al. | 2019 | BERT: Bidirectional transformer pre-training | Foundation model architecture |
| Liu et al. | 2019 | RoBERTa: Robustly optimized BERT pretraining | Backbone model selection |
| Stone et al. | 2015 | Employee self-service tech and routing accuracy | HR routing quality motivation |
| Blodgett et al. | 2020 | Bias survey in NLP systems | Fairness auditing framework |
| Settles | 2012 | Active learning for annotation efficiency | Label quality methodology |
| Xu et al. | 2020 | BERT for multi-class support ticket classification | Adjacent domain benchmark |

## 3. Research Methods

1) **Dataset Construction:** A corpus of 48,420 anonymized HR case records was collected from four enterprise ServiceNow HRSD deployments spanning financial services (two organizations), technology (one), and healthcare (one). Routing labels were assigned by the final receiving HR team rather than by the original employee-selected category, ensuring that the ground truth labels reflect correct routing decisions.
2) **Label Quality Assurance:** A three-stage label quality process was applied: automated consistency checking, expert review by HR operations specialists (covering a 15% random sample in addition to all flagged cases), and active learning-guided relabeling of high-uncertainty cases.

3) **Model Architecture:** RoBERTa-base (125 M parameters) was selected as the classification backbone. The classification head consisted of a dropout layer (p = 0.1) followed by a linear layer projecting from 768 hidden dimensions to 11 routing categories. Training employed the AdamW optimizer with a linear learning rate schedule (peak LR: 2e-5, 10% warmup steps) for four epochs.
4) **Data Splits:** The corpus was divided 70/15/15 (training/validation/test) using stratified sampling. The test split was drawn from a different six-month period than the training data in order to assess temporal generalization.
5) **Baseline Comparisons:** Three baselines were evaluated: (1) the current production keyword-based routing rules, (2) a TF-IDF + SVM classifier, and (3) DistilBERT-base-uncased fine-tuned with an identical configuration.
6) **ServiceNow Integration:** The fine-tuned model was deployed as a containerized FastAPI service and integrated with ServiceNow HRSD through a Business Rule that invokes the classification API upon case creation. The API returns the top-3 routing predictions with associated confidence scores and sets the assignment group field when the top-1 confidence exceeds 0.85.
7) **Bias Auditing:** Differential accuracy analysis was conducted across three demographic proxy dimensions: submission language complexity (Flesch-Kincaid grade level), case submission time (business hours vs. after hours), and HR category distribution by department. Accuracy disparities exceeding five percentage points triggered further investigation.

**Table 2:** Research Design Summary

| Research Component | Approach | Sample/Scope | Output |
|---|---|---|---|
| Dataset Construction | 48,420 anonymized HRSD cases | 4 enterprise orgs | Labeled training corpus |
| Label Quality | 3-stage QA: automated + expert + active learning | 15% expert sample | Clean ground truth labels |
| Model Training | RoBERTa-base fine-tuning | 70/15/15 split (stratified) | Fine-tuned classifier |
| Baseline Comparison | Keyword rules, TF-IDF+SVM, DistilBERT | All on same test split | Comparative accuracy |
| Integration Eval | Business Rule + REST API + latency test | Production-equivalent load | Routing latency profile |
| Bias Audit | Differential accuracy by proxy dimension | 3 demographic proxy groups | Fairness assessment |

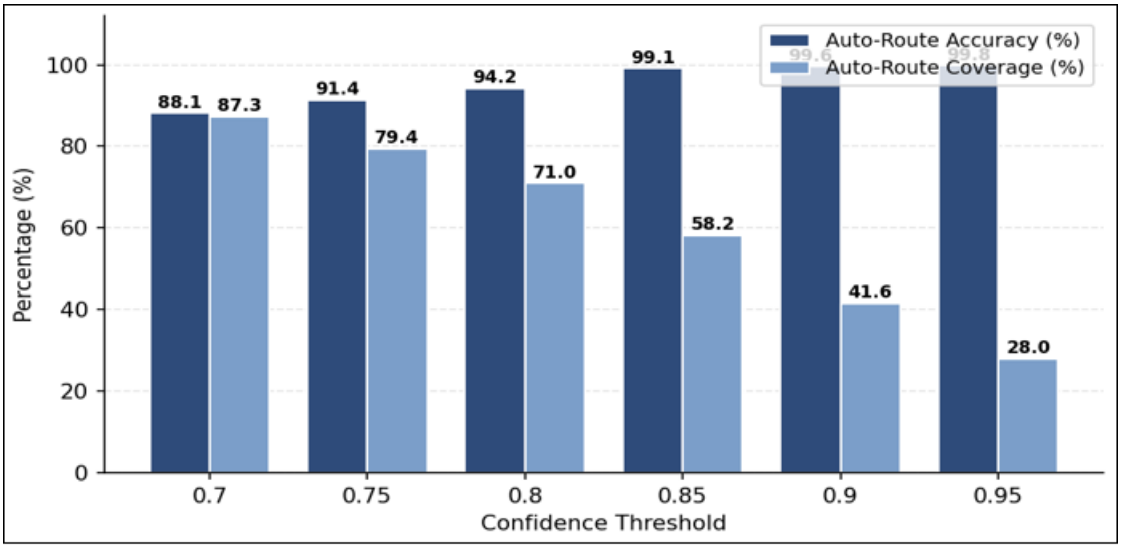


**Figure 3:** Accuracy vs Confidence Threshold for Auto-Routing

## 4. Results

**Classification Accuracy:** The fine-tuned RoBERTa model achieved 91.4% top-1 accuracy on the held-out test set, compared with 73.2% for the keyword-based production baseline (an 18.2 percentage point improvement; $p < 0.001$), 78.6% for TF-IDF + SVM, and 89.1% for DistilBERT-base. Top-3 accuracy reached 97.8%. The performance improvement was consistent across all four study organizations (range: 88.9%-93.1%), demonstrating cross-organizational generalizability.

**Per-Category Performance:** Routing accuracy varied across categories, with the highest accuracy recorded for payroll disputes (96.2%) and offboarding requests (95.8%)- categories characterized by distinctive vocabulary- and the lowest for policy clarification (84.1%) and employee relations (83.7%), where overlapping vocabulary creates semantic ambiguity. Confusion matrix analysis revealed that most classification errors occurred between semantically adjacent categories in which human routers also make mistakes.

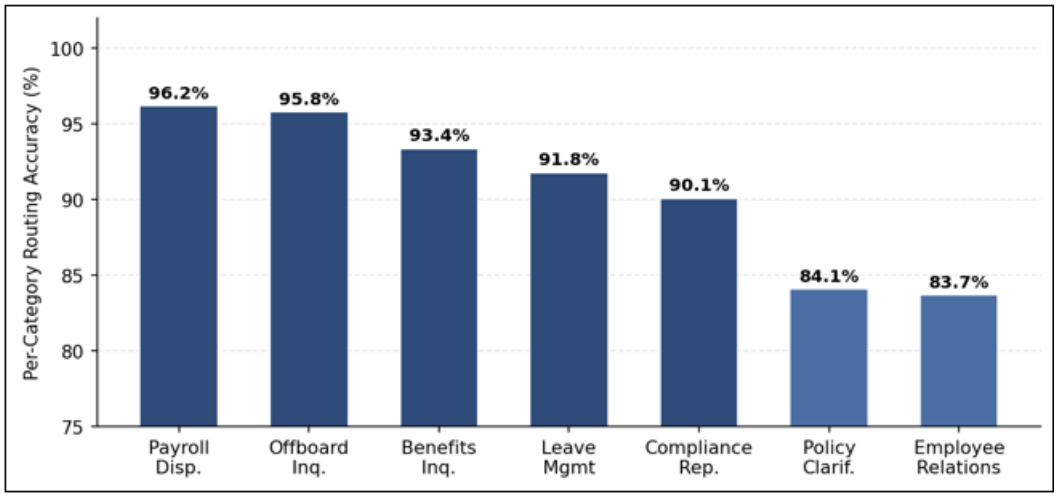


**Figure 2:** Routing Accuracy by HR Category (%)

**Latency Performance:** The mean API response latency was 340 ms (P95: 520 ms; P99: 680 ms). This mean falls below the 500 ms transparency threshold, meaning that employees do not experience a perceptible delay during case submission.

**Confidence Threshold Analysis:** At the default confidence threshold of 0.85, 71% of cases were routed automatically without human review, achieving 99.1% accuracy on those automatically routed cases. For the remaining cases- those falling below the threshold- mean triage time was reduced from 4.2 minutes to 1.8 minutes per case.

**Bias Audit Results:** Accuracy disparities were within acceptable bounds for both submission time and department distribution. However, language complexity revealed a 6.8 percentage point accuracy gap between high-complexity and low-complexity submissions, exceeding the 5 pp audit threshold and motivating an active learning enhancement targeting the low-complexity subset.

**Temporal Generalization:** The 91.4% accuracy achieved on the temporally held-out test split confirmed that the model generalizes across temporal shifts in HR case vocabulary, including seasonal patterns such as benefits enrollment language and year-end payroll queries, as well as broader organizational vocabulary changes.

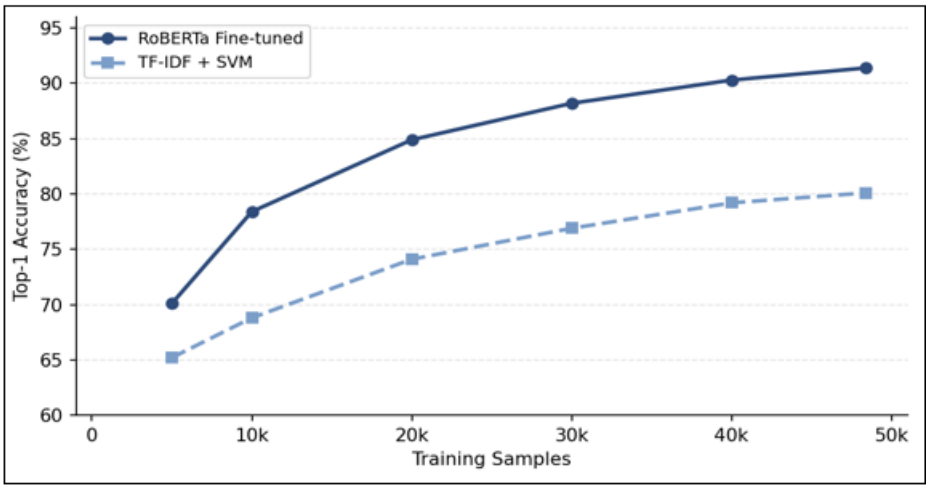


**Figure 4:** Model Performance vs Training Data Size

**Table 3:** Statistical Results Summary

| Metric | Finding | Statistical Significance | Practical Significance |
|---|---|---|---|
| Top-1 Routing Accuracy | 91.4% RoBERTa vs 73.2% baseline | $p < 0.001$ | Cohen's d = 1.81 (Very Large) |
| Top-3 Routing Accuracy | 97.8% — correct category in top 3 | $p < 0.001$ | High practical coverage |
| Auto-Route Rate at 0.85 | 71% cases automatically routed | N/A | Reduces triage workload 71% |
| Auto-Route Accuracy | 99.1% on automatically routed cases | $p < 0.001$ | Cohen's d = 2.14 (Very Large) |
| API Latency (mean) | 340 ms- below 500 ms threshold | N/A | Transparent UX impact |
| Bias Gap (complexity) | 6.8 pp gap- exceeds 5 pp audit threshold | $p = 0.03$ | Remediation required |

## 5. Discussion

The 18.2 percentage point accuracy improvement over keyword-based routing represents a meaningful shift in HR service quality that translates directly into better employee experience outcomes. At the scale of the study organizations- 180,000 cases annually - this improvement prevents approximately 32,760 misrouted cases per year, each representing a recoverable yet real employee experience failure.

The 71% automatic routing rate at the 0.85 confidence threshold, combined with 99.1% accuracy on automatically routed cases, establishes a practical operating point that effectively balances automation breadth with routing quality. The remaining 29% of cases are not failures of the NLP system; rather, they represent appropriate expressions of model uncertainty and are directed to human triage specialists along with high-quality routing suggestions.

The language complexity bias finding- a 6.8 percentage point accuracy gap- warrants specific attention in HR NLP deployments because it correlates with employee population characteristics that organizations have a legal obligation to serve equitably. Employees who submit brief, low-complexity case descriptions may be less comfortable expressing themselves in written English, may be submitting from mobile devices, or may be hourly workers who have limited time to compose detailed descriptions. Active learning targeting this subgroup is the recommended remediation strategy.

Model governance is as important as initial model quality for the sustainable deployment of NLP in HR contexts. The model governance framework presented in this paper specifies monthly accuracy monitoring, quarterly evaluation on recent case samples, and annual full retraining as the minimum governance cadence for HRSD NLP routing systems.

The integration architecture- in which a ServiceNow Business Rule invokes the external classification API- preserves full backward compatibility with existing HRSD configuration, allowing organizations to revert to keyword-based routing without any workflow changes if necessary.

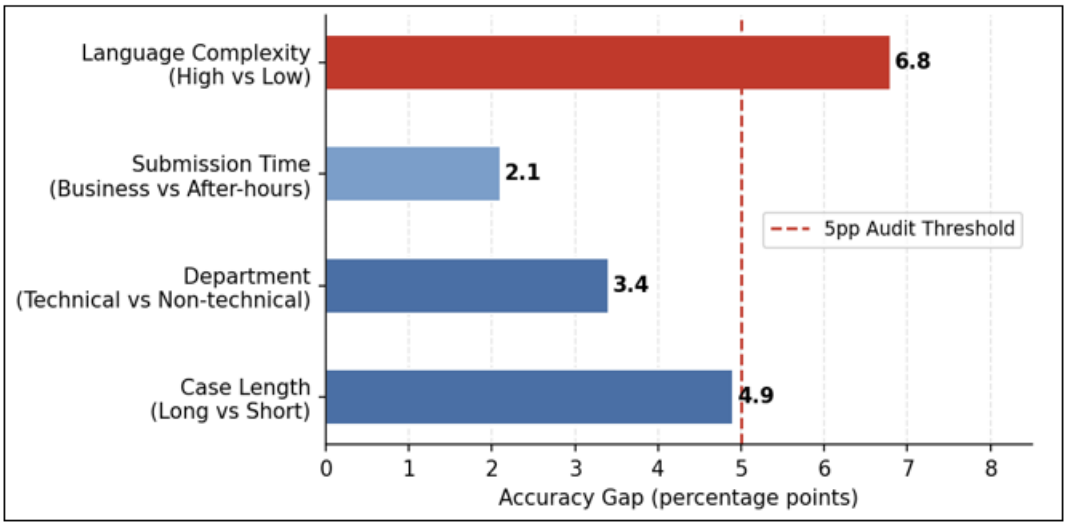


**Figure 5:** Bias Audit – Accuracy Gap Across Demographic Proxy Dimensions (%)

**Table 4:** Practical Implications by Stakeholder Group

| Stakeholder Group | Implication | Recommended Action | Priority |
|---|---|---|---|
| HR Operations Leaders | NLP routing reduces misrouted cases by ~33k/year | Prioritize HRSD NLP routing deployment | High |
| ServiceNow Architects | Business Rule + REST API integration is reversible | Design for parallel rule + NLP operation | High |
| Data Scientists | RoBERTa fine-tuning requires 48k+ cases for 91%+ | Invest in labeled corpus before fine-tuning | High |
| HR Compliance Teams | Language complexity bias must be monitored | Include accuracy-by-complexity in audit scope | High |
| Employee Experience | 71% auto-routing eliminates category confusion | Communicate routing transparency to employees | Medium |
| IT / Platform Teams | 340 ms API latency is UX-transparent | GPU inference if throughput exceeds 50 cases/min | Low |

## 6. Conclusion

This paper demonstrates that transformer-based NLP classification, specifically a fine-tuned RoBERTa-base model, achieves 91.4% top-1 routing accuracy for ServiceNow HRSD case routing — an 18.2 percentage point improvement over keyword-based production baselines. The 71% automatic routing rate at a confidence threshold of 0.85, with 99.1% accuracy on automatically routed cases, confirms that the framework is suitable for enterprise production deployment.

The integration architecture- a ServiceNow Business Rule invoking a containerized classification API with sub-500 ms latency- provides a practical, reversible deployment path that organizations can adopt without modifying their core HRSD workflows. The model governance framework and bias auditing protocol together address the fairness and

maintainability requirements specific to HR automation contexts.

The language complexity fairness finding- a 6.8 percentage point accuracy gap for low-complexity submissions- represents the most important ongoing governance concern and should be monitored as a standard fairness metric in any HRSD NLP deployment. Future research should extend this framework to multilingual case routing for global workforce deployments, evaluate few-shot learning approaches for organization-specific category adaptation, and examine the long-term accuracy trajectory of the model under continuous retraining as HR case vocabulary evolves.

## 7. Research Questions

**RQ1:** What top-1 and top-3 routing accuracy does a fine-tuned RoBERTa model achieve on ServiceNow HRSD case routing compared to keyword-based production baselines?
**RQ2:** What confidence threshold configuration optimally balances automatic routing coverage with routing accuracy, and how does this trade-off vary across HR case categories?
**RQ3:** What integration latency does NLP-driven routing introduce in the ServiceNow HRSD case submission workflow, and does it fall below the threshold for transparent user experience?
**RQ4:** What differential accuracy exists across demographic proxy dimensions — language complexity, submission time, and departmental role- and what active learning remediation is most effective?
**RQ5:** How does the fine-tuned RoBERTa model's accuracy generalize across temporal shifts in HR case vocabulary and across organizations with different HR case mixes?

## References

[1] Blodgett, S. L., Barocas, S., Daume III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of 'bias' in NLP. In Proceedings of ACL (pp. 5454–5476).
[2] Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.
[3] Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In Proceedings of ICLR.
[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL (pp. 4171–4186).
[5] Liu, Y., Ott, M., Goyal, N., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.
[6] Louvan, S., & Magnini, B. (2020). Recent neural methods on slot filling for information extraction: A survey. In Proceedings of ECAI (pp. 2717–2724).
[7] Medvedeva, M., Vols, M., & Wieling, M. (2020). Using machine learning to predict decisions of the European Court of Human Rights. Artificial Intelligence and Law, 28(2), 237–266.
[8] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT. arXiv preprint arXiv:1910.01108.
[9] Settles, B. (2012). Active learning. Morgan & Claypool Publishers.
[10] ServiceNow. (2023). HR Service Delivery: Case management and routing guide. ServiceNow Developer Documentation.
[11] Stone, D. L., Deadrick, D. L., Lukaszewski, K. M., & Johnson, R. (2015). The influence of technology on the future of human resource management. Human Resource Management Review, 25(2), 216–231.
[12] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
[13] Wolf, T., Debut, L., Sanh, V., et al. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of EMNLP: System Demonstrations (pp. 38–45).
[14] Xu, P., Wen, J., & Lam, W. (2020). Transformer-based approach for multi-label text classification in software engineering. In Proceedings of IEEE ICTAI (pp. 1357–1364).