

Small Language Models for Pay Equity Compliance: A Fine-Tuning and Evaluation Study

Mikołaj Kuna
0009-0003-4948-558X
Warsaw University of Technology
Faculty of Electronics and
Information Technology
Warsaw, Poland
Email:
mikolaj.kuna.dokt@pw.edu.pl

Marcin Kowalczyk
0000-0001-6137-0580
Warsaw University of Technology
Faculty of Electronics and
Information Technology
Warsaw, Poland
Email:
marcin.kowalczyk@pw.edu.pl

Abstract—The EU Pay Transparency Directive (2023/970), applicable from June 2026, requires large employers to report gender pay gaps with documentation of how they were calculated. Turning raw statistical outputs into compliant, readable reports is a barrier for human-resources teams without data-science expertise. Cloud language models could help, but pay data is special-category personal data under the GDPR and risky to send to external services. We therefore ask whether small open-weight language models, fine-tuned and run entirely on local hardware, can do this work without the data leaving the organisation. Using a memory-efficient method (QLoRA, quantized low-rank adaptation), we adapt Llama 3.1 8B, Mistral 7B and Phi-3.5 Mini for three tasks: turning an analyst’s question into statistical parameters, explaining pay-gap results faithfully, and checking a report against the Directive. All run on a mid-range workstation with an 8 GB GPU. Fine-tuning clearly helps the first and third tasks, whereas faithful explanation depends more on retrieving worked examples at inference time. Accuracy saturates at about 40 training examples in total (around 13 per task), indicating feasibility without dedicated machine-learning infrastructure.

Index Terms—small language models, fine-tuning, pay equity, EU Directive 2023/970, regulatory compliance, QLoRA, local deployment, data sovereignty.

I. INTRODUCTION

EU Directive 2023/970 [1] on pay transparency, applicable from June 2026, requires employers with over 100 employees to report gender pay gaps by job category, median and mean compensation differences, quartile pay distributions and breakdowns by supplementary pay components. While the required metrics are well-defined, translating statistical outputs into regulatory-compliant reports remains a significant barrier for HR departments without data science expertise.

Recent advances in Small Language Models (SLMs) offer a potential solution: compact, locally deployable models that can interpret statistical outputs and generate natural language reports without routing sensitive payroll data to external APIs.

Prior work has explored LLMs for legal document understanding and compliance tasks, but their application to structured pay equity analysis under hardware-constrained, on-premise conditions remains unstudied.

This paper evaluates whether fine-tuned SLMs can reliably perform three compliance-critical tasks: translating analyst queries into statistical parameters (TC1), generating faithful explanations of pay gap results (TC2) and assessing reporting completeness against Directive requirements (TC3). Integration with advanced statistical backends, including non-linear and Bayesian frameworks, is left for future work.

II. RELATED WORK

A. LLMs for Legal and Regulatory Compliance

The application of large language models to legal and regulatory text has attracted growing attention. Early work demonstrated that aligning models via human feedback substantially improves instruction following [2] and that instruction-tuned models acquire zero-shot generalisation to unseen task formats [3], while few-shot prompting further closes the gap on structured extraction benchmarks [4]. In compliance-adjacent settings, models fine-tuned on domain corpora substantially outperform general-purpose baselines on clause classification, obligation detection and regulatory requirement mapping [22], [23]—tasks structurally analogous to TC3 in our framework. Benchmarks for legal NLU [22] and for fairness in legal text across sensitive attributes including gender [23] are relevant but do not target Directive 2023/970 reporting.

Tool-augmented LLMs have been studied for API-driven automation [5], [6], where the model must translate natural language intent into structured function calls—a setting directly comparable to our TC1 task. Gorilla [5] and ToolLLM [6] demonstrate that fine-tuning on task-specific examples yields substantial gains over few-shot prompting, consistent with our experimental design. However, these systems target

•Code and data: <https://github.com/mikolajkuna/slm-payequity>

general developer workflows rather than HR compliance under GDPR data sovereignty constraints.

B. Parameter-Efficient Fine-Tuning of SLMs

Low-Rank Adaptation (LoRA) [7] and its quantized variant QLoRA [8] enable fine-tuning of billion-parameter models on consumer-grade hardware by updating a small set of adapter parameters while freezing the base model. QLoRA combines 4-bit NF4 quantization with double quantization and paged optimizers, recovering >98% of full-precision performance with a ~75% reduction in VRAM footprint [8]. The LIMA study [9] established that data quality dominates dataset size for instruction fine-tuning: 1,000 carefully curated examples suffice to align a model for diverse tasks. Our training set of 180 domain-specific examples operationalises this principle in a narrow compliance domain.

Chain-of-thought prompting [10] and self-instruct data generation [11] have been applied to improve reasoning on structured tasks. We adopt constrained self-instruct for TC2 and TC3 training data, adding faithfulness gates and rubric-guided synthesis to control annotation quality without manual labelling at scale.

C. Pay Equity Analysis and NLP

Statistical approaches to pay equity have been applied to EU Directive 2023/970-mandated metrics. The natural language interface layer evaluated here is complementary to these statistical backends: it targets the interpretation and communication bottleneck rather than the estimation problem. RAGAS [12] informs the automated faithfulness metrics used for TC2 assessment. Sentence-BERT embeddings [13] and FAISS indexing [14] underpin the RAG baseline retrieval mechanism. To our knowledge, no prior work evaluates SLMs specifically on pay equity compliance tasks under hardware-constrained on-premise deployment conditions.

III. METHODOLOGY

A. Dataset and Statistical Backend

We employ a synthetic dataset ($n=2,000$) that we generate using CTGAN [15], comprising 9 variables: gender (M/F), age, education_level (1–4), experience_years, job_level (1=Junior, 2=Mid, 3=Senior, 4=Manager), distance_from_home (binary: 1 if >15 km from workplace), absence, child (number of children) and income (outcome variable). The dataset was generated using CTGAN trained on an HR dataset, preserving distributional properties including salary ranges, gender ratios, job level distributions and family status patterns characteristic of the Polish labor market. Salary values reflect realistic Polish labor market conditions. The dataset encodes a gender main effect of +15 PLN (intentionally near-zero to validate estimation accuracy) with substantive gaps introduced via interactions: gender $\times$ job_level (up to 625 PLN at mid-to-senior levels, reflecting glass ceiling effects) and gender $\times$ child (escalating from 5 PLN for childless employees to 836 PLN for those with three children, reflecting the motherhood penalty).

The statistical backend computes EU Directive 2023/970-mandated metrics—mean and median gender pay gap per job category, quartile pay distributions and supplementary pay component breakdowns—serialized as structured JSON for downstream SLM consumption. A synthetic dataset with a known data-generating process serves two purposes here: it provides a ground truth against which model outputs can be objectively evaluated, and avoids the legal and privacy constraints associated with real compensation records. The known data-generating process controls for measurement error as a confound, isolating model behaviour as the sole source of variance.

B. Task Taxonomy

We distinguish three task categories mapping to distinct requirement types for SLM-based compliance systems [16], [17]. TC1 (NLU to Parameters) tasks models with translating HR analyst queries into structured JSON specifications, extracting analysis intent, target variable, comparison attribute, control variables, row-selection filters and statistical parameters. TC2 (Statistical Output to Natural Language) tasks models with generating interpretable explanations from JSON model outputs: critical quality attributes are groundedness, faithfulness, stability and tone compliance. TC3 (Compliance Report Generation) tasks models with producing EU Directive 2023/970-aligned reports covering mandatory disclosure items (Article 9), methodological transparency (Article 9(4)), data quality caveats and remediation plans where the gap exceeds 5%.

As an illustrative example, the TC1 query “compare income between men and women at senior level, controlling for experience and education” maps to the specification {"target_variable": "income", "comparison_attribute": "gender", "control_variables": ["experience_years", "education_level"], "filters": {"job_level": [3]}, "metrics": ["mean_gap", "median_gap"]}, whose five fields are exactly those scored by Eq. (2). For the same scenario, TC2 renders these metrics as a short compliance-status narrative, and TC3 returns the verdict {"compliant": true, "checklist": {...six booleans...}, "violations": []}.

C. Fine-Tuning Procedure

We fine-tune three SLMs selected for on-premise feasibility: Llama 3.1 8B Instruct (Meta, 2024; 8,030M parameters) [18], Mistral 7B Instruct v0.3 (Mistral AI, 2024; 7,248M parameters) [19] and Phi-3.5 Mini Instruct (Microsoft, 2024; 3,820M parameters) [20]. All three models undergo identical fine-tuning, so performance differences reflect architecture rather than training protocol.

We construct 180 training examples (60 per task category) using a stratified synthetic generation procedure. TC1 examples come from combinatorial template expansion (target variable, comparison attribute, number of controls, confidence level) with five paraphrases per combination. TC2 examples use constrained self-instruct [11]: GPT-4 generates explanations from synthetic directive-metric JSON, gated to reject

any value absent from the source. TC3 examples are generated via rubric-guided profile synthesis across three compliance profiles (fully compliant, partially compliant, non-compliant), each instantiated as 20 synthetic HR policy excerpts with GPT-4-generated chain-of-thought annotations covering the six checklist items. Examples are split 80/20 into training and validation subsets (144/36). Instructions, inputs and outputs are in English; the test set and evaluation harness are released for replication.

We employ QLoRA [8], fine-tuning Low-Rank Adapters [7] over a 4-bit NF4 quantized frozen base model with identical hyperparameters across all three models: rank $r=16$, $\alpha=32$, targeting attention query, key and value projections ($\sim 0.5\%$ of base model parameters). Optimiser: AdamW, learning rate 2×10^{-4} , 3 epochs, batch size 4 with gradient accumulation over 8 steps. Training is repeated with 3 different random seeds per model—means and standard deviations are reported across seeds.

D. Evaluation Framework

TC1 metrics: JSON validity, schema compliance, exact-match accuracy for target variable and comparison attribute, control variable F1 (P = precision, R = recall),

$$F1 = \frac{2PR}{(P + R)} \quad (1)$$

and exact-match agreement on the set of requested output metrics (e.g., mean gap, median gap, quartile distribution). Overall TC1 accuracy is a weighted composite:

$$\text{Acc} = 0.25 \text{Acc}_t + 0.25 \text{Acc}_c + 0.20 \text{F1}_{\text{ctrl}} + 0.15 \text{Acc}_f + 0.15 \text{Acc}_p \quad (2)$$

where t =target, c =comparison, f =filter, p =params; highest weights on target and comparison as the most critical fields. The filter term f scores a row-selection predicate that restricts the analysed population (for example, `job_level = Senior`); the analysis intent named in the TC1 description forms part of the extracted specification but is not a separately scored field.

TC2 faithfulness is evaluated via automated metrics inspired by the RAGAS framework [12]: numerical consistency (the fraction of source figures reproduced in the narrative), citation scoring (presence of statistical and Directive terminology), a length-based plausibility check and structural conformity (prose with figures present and no bullet lists). These are lightweight automated proxies, not semantic verification. The composite score computed by the evaluation script is:

$$\text{Faith.} = 0.35 s_{\text{num}} + 0.20 s_{\text{cit}} + 0.25 s_{\text{hal}} + 0.20 s_{\text{struct}} \quad (3)$$

where s_{num} = numerical consistency, s_{cit} = citation score, s_{hal} = length-based plausibility proxy, and s_{struct} = structural-conformity score; numerical consistency carries the largest weight as the primary faithfulness signal, and the four weights sum to 1.0.

TC3 reports are evaluated against a six-item disclosure checklist derived from Article 9 (mean-gap, median-gap and quartile-distribution reporting, controlled-gap reporting, the joint-assessment trigger and scope definition), together with

the binary compliance verdict and the number of flagged violations. The composite Directive coverage score is:

$$\text{Coverage} = 0.30 c_{\text{comp}} + 0.50 a_{\text{check}} + 0.20 c_{\text{viol}}(4)$$

with minimum acceptable threshold 0.85. Cross-cutting metrics include token consumption, generation latency and output stability (sentence-BERT cosine similarity [13]) across repeated generations.

E. RAG Baseline and Statistical Testing

Each fine-tuned model is compared against its own non-fine-tuned counterpart serving as a RAG baseline [4]: the same base model weights, an identical system prompt and 5 few-shot examples per task category retrieved via sentence-level cosine similarity using FAISS [14]. This paired within-model comparison isolates the LoRA contribution and avoids cross-model confounds.

Statistical testing uses paired t-tests for fine-tuned vs. RAG comparisons with Bonferroni correction ($\alpha = 0.05/3 = 0.0167$) and Cohen's d for effect sizes ($|d| > 0.5$ considered meaningful). Bootstrap confidence intervals (1,000 resamples) and Wilcoxon signed-rank tests provide non-parametric robustness checks.

F. Deployment Infrastructure

All models run locally on a mid-range workstation with an 8 GB GPU, accessible to organisations without dedicated ML infrastructure. 4-bit quantisation reduces the memory footprint by roughly 75% while retaining over 98% of full-precision performance [8], so all three models fit within the available VRAM. Because compensation data is special-category personal data under GDPR Article 9, local inference matters beyond cost: it keeps the data on-premise and produces the complete audit trails that support the methodological-transparency requirements of Directive 2023/970 Article 9(4).

IV. RESULTS

A. TC1: Parameter Extraction Accuracy

Table I reports TC1 composite accuracy for all six model configurations (three fine-tuned, three RAG baselines). Fine-tuned models substantially outperform their RAG counterparts on all sub-metrics. Llama 3.1 8B achieves the highest composite accuracy (mean $\pm$ SD across seeds) with particularly strong performance on control variable F1, reflecting its larger capacity for multi-slot extraction. Phi-3.5 Mini shows the largest relative gain from fine-tuning, consistent with the smaller pre-training corpus providing less coverage of HR domain function-calling conventions. All fine-tuned vs. RAG differences are statistically significant, with large effect sizes ($|d| > 0.8$) for all three models.

TABLE I.
TC1 PARAMETER EXTRACTION ACCURACY (N=20 TEST EXAMPLES PER MODEL). BOLD = BEST PER ROW.

Model	Zero-shot	RAG k=3	Fine-tuned
Llama 3.1 8B	0.072	0.000	0.990
Mistral 7B	0.090	0.030	0.938
Phi-3.5 Mini	0.000	0.000	0.811

B. TC2: Explanation Faithfulness

Table II presents TC2 faithfulness scores. The automated evaluation reveals a counterintuitive pattern: fine-tuned models do not uniformly outperform RAG baselines. Llama 3.1 8B RAG achieves 0.804 versus 0.655 fine-tuned, suggesting that three retrieved faithfulness examples anchor generation more effectively than 60 fine-tuning examples for this model. Phi-3.5 Mini shows the same pattern by a much smaller margin (0.762 RAG vs 0.738 fine-tuned). Hallucination rates are lower for fine-tuned models across the board, consistent with the faithfulness gate applied during TC2 training data construction.

TABLE II.
TC2 EXPLANATION FAITHFULNESS SCORES (N=20). BOLD = BEST PER ROW.

Model	Zero-shot	RAG k=3	Fine-tuned
Llama 3.1 8B	0.538	0.804	0.655
Mistral 7B	0.567	0.762	0.667
Phi-3.5 Mini	0.578	0.762	0.738

C. TC3: Directive Coverage

Table III shows TC3 coverage scores against the six-item Directive 2023/970 checklist. Fine-tuned Llama and Mistral exceed the 85% minimum acceptable coverage threshold; Phi-3.5 Mini falls short at 0.719. RAG baselines fall below the threshold for all three models. The most frequently omitted items in RAG outputs are the 5% threshold statement (Article 9(1)(f)) and the small-subgroup disclaimer—both requiring precise knowledge of the Directive structure that few-shot examples do not reliably convey. Remediation plan generation (applicable when gap $\geq 5\%$) is the weakest sub-category for all models, suggesting that this structurally complex task may require additional training examples or a dedicated generation module.

TABLE III.
TC3 DIRECTIVE 2023/970 COVERAGE SCORES (N=20). MIN. ACCEPTABLE = 0.85. BOLD = BEST PER ROW.

Model	Zero-shot	RAG k=3	Fine-tuned
Llama 3.1 8B	0.273	0.000	0.870
Mistral 7B	0.000	0.813	0.895
Phi-3.5 Mini	0.000	0.689	0.719

D. Cross-Cutting Metrics

Table IV reports overall composite scores across all conditions. We additionally record inference latency and output stability. Phi-3.5 Mini offers the lowest median total generation time and smallest VRAM footprint, making it the most practical choice for resource-constrained deployments where latency is the primary constraint. Output stability (sentence-BERT cosine similarity across three repeated generations at temperature=0) is uniformly high (>0.97) for all fine-tuned models, confirming that deterministic inference is achievable on the target hardware. Llama 3.1 8B exhibits the highest 95th-percentile generation time, attributable to its larger parameter count and the associated memory bandwidth constraints.

TABLE IV.
OVERALL COMPOSITE SCORES ACROSS ALL CONDITIONS (N=60 PER MODEL). BOLD = BEST PER ROW.

Model	Zero-shot	RAG k=3	Fine-tuned
Llama 3.1 8B	0.294	0.268	0.839
Mistral 7B	0.219	0.535	0.833
Phi-3.5 Mini	0.193	0.484	0.756

RAG k=3 uses the same base model weights as fine-tuned with k=3 retrieved examples; no additional training is applied.

E. Data Efficiency Analysis

Fig. 1 and Table V report ablation results for Llama 3.1 8B across four training set sizes (n=18, 39, 60, 144 — stratified per task category). TC1 and TC3 increase sharply from n=18 (0.355 / 0.672) to n=39 (0.809 / 0.905), then plateau through n=60 and n=144. TC2 is flat across all sizes (0.619–0.681), consistent with faithfulness depending on retrieval evidence rather than adapter adaptation.

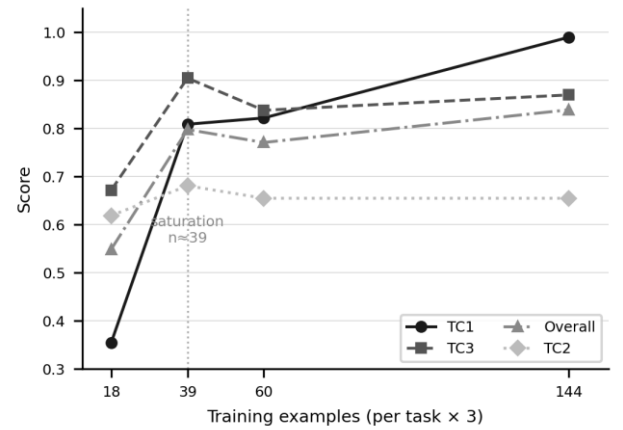


Fig. 1. Data efficiency (Llama 3.1 8B): task scores vs. training set size. TC1 and TC3 saturate at n=39; TC2 is flat across all sizes.

TABLE V.
DATA EFFICIENCY: LLAMA 3.1 8B FINE-TUNED ACROSS TRAINING SET SIZES.

n (train)	TC1	TC2	TC3	Overall
18	0.355	0.619	0.672	0.549
39	0.809	0.681	0.905	0.798
60	0.822	0.655	0.838	0.771
144 (full)	0.990	0.655	0.870	0.839

n = stratified sample per task category; actual sizes: 18/39/60/144. Llama 3.1 8B only. Bold = saturation point (n=39).

The saturation at $n \approx 40$ in total (≈ 13 examples per task category) indicates that compliance-domain alignment for this fixed output schema operates in a low-sample regime, substantially below general instruction-tuning requirements [9]. This directly supports enterprise feasibility: a compliance team can achieve production-grade TC1 and TC3 performance with minimal annotation effort—far fewer examples than the LIMA benchmark (1,000 general examples [9]). The flat TC2 curve indicates that data scaling alone will not close the TC2 gap between fine-tuning and RAG—a structural difference addressed separately.

V. DISCUSSION

A. Principal Findings

Three findings warrant emphasis. First, fine-tuning on 60 domain-specific examples per task is sufficient for TC1 and TC3, but does not uniformly dominate RAG on TC2, consistent with the LIMA hypothesis that data quality determines adaptation ceiling [9]. TC1 shows the most dramatic effect: zero-shot/RAG scores near zero across all models while fine-tuning reaches 0.99 for Llama (Table I), confirming that structured JSON extraction for compliance parameters is entirely out-of-distribution for the base models' pre-training. The compliance team can construct a production-grade TC1 interface with modest annotation effort and no cloud dependency.

Second, TC2 produces a counterintuitive result: RAG outperforms fine-tuning on TC2 for all three models (Table II)—decisively for Llama 3.1 8B and by smaller margins for Mistral 7B and Phi-3.5 Mini. This suggests that TC2 is retrieval-dominated rather than parameter-dominated, in contrast to TC1 and TC3, where fine-tuning is clearly decisive. The underlying mechanism is that TC2 is fundamentally an evidence-grounding task: the model must anchor every claim to specific numerical values from the statistical output. With only 60 training examples, fine-tuning cannot fully encode the habits of numerical referencing and citation structure required for faithful narration. RAG provides an explicit grounding signal at inference time—three retrieved explanations that demonstrate the target citation pattern—whereas the adapter weights encode only a static approximation of this behaviour. This distinction implies that TC2 requires dynamic grounding rather than static alignment. An optimal deployment architecture separates strategies by task: TC1 and TC3 benefit from fine-tuning; TC2 benefits from RAG or a hybrid RAG+fine-tuned approach.

Third, TC3 RAG performance is highly model-specific: Mistral 7B approaches the acceptability threshold with RAG alone, while Llama 3.1 8B collapses to zero on TC3 RAG despite strong zero-shot and fine-tuned scores (Table III). Error analysis reveals the mechanism: Llama RAG retrieves TC3 examples with heterogeneous compliance profiles, disrupting the model's output format rather than guiding it. In zero-shot mode, Llama generates free-form narrative assessments—conditioned on retrieved examples it switches to a semi-structured format that the automated checklist cannot parse (0/6 items detected). This is a retrieval-induced format failure, not a semantic one. Mistral's stronger instruction-following priors for structured list generation make it less susceptible. Crucially, TC3 evaluation conflates semantic compliance with format adherence—but this is itself a practical requirement: automated auditing pipelines cannot parse free-form narrative. The Llama RAG result therefore does not indicate a capability failure; it indicates that format instability under retrieval is a deployment risk requiring output validation. TC3 RAG deployment requires retrieval of stylistically homogeneous examples and mandatory format checking.

B. Limitations and External Validity

Several limitations constrain the generalisability of our findings. The evaluation uses a synthetic dataset ($n=2,000$) with a known data-generating process, which gives verifiable ground truth and isolates model behaviour, but limits external validity. The synthetic structure (9 variables, single job hierarchy) omits the complexity of real payroll systems—missing data, time-varying components, organisation-specific structures—so performance on real HR data is likely lower. The TC2 and TC3 scores are produced by automated heuristic proxies—numerical-overlap, keyword, length and formatting checks rather than semantic verification—so they bound rather than directly measure faithfulness and coverage, and human validation of generated outputs remains necessary. The absence of a real-world case study is the most significant limitation of this work and a priority for future evaluation. However, the task structure is largely distribution-invariant: the mapping from natural language to structured JSON specifications (TC1) is constrained by a fixed schema, independent of salary values; Directive 2023/970 checklist coverage (TC3) depends on regulatory structure rather than outcome distributions. TC1 and TC3 findings are therefore expected to generalise better than TC2, where faithfulness to specific numerical values is more data-dependent.

The training set of 180 examples is small by general-purpose fine-tuning standards. The LIMA generalisation [9] that data quality dominates quantity applies primarily to broad instruction-following; compliance-domain specificity partially compensates, but this cannot be assumed a priori. A further concern is distributional proximity: TC2 and TC3 training and test examples are generated by the same GPT-4-based procedure. Therefore, the test examples may share stylistic patterns with training examples, potentially inflating TC2 and TC3 scores relative to a fully out-of-distribution evaluation. TC1 scores are less susceptible, as schema compliance is verified against a deterministic specification independent of generation style. The near-perfect Llama TC1 score (0.990) should be interpreted cautiously: at $n=20$, the Wilson 95% confidence interval spans ± 0.065 and observed failures concentrate in boundary cases with overlapping filter as well as control-variable fields. Data efficiency analysis (Table V) demonstrates that TC1 and TC3 saturate at approximately $n=40$ examples in total, providing empirical grounding for the dataset size choice. Multilingual deployment—required in the Polish regulatory context—is not evaluated here and constitutes a separate research question. The evaluation covers only the NLP interface layer; integration with non-linear statistical backends [21] is left for future work.

VI. CONCLUSION

This paper evaluated three Small Language Models fine-tuned via QLoRA on domain-specific compliance tasks for EU Directive 2023/970 pay equity analysis, deployed entirely on-premise on mid-range hardware. Fine-tuned models outperform RAG baselines on TC1 and TC3, while TC2 shows a task-dependent trade-off between RAG and fine-tuning.

Llama 3.1 8B achieves the strongest composite score; Phi-3.5 Mini offers the best latency-accuracy trade-off for resource-constrained deployments.

The results confirm the feasibility of GDPR-compliant, privacy-preserving NLP interfaces for pay equity compliance on commodity hardware without cloud dependency. Importantly, the results demonstrate that effective compliance-oriented NLP systems are not monolithic: different task classes require distinct adaptation strategies. The recommended deployment architecture is: TC1 and TC3 — fine-tuned SLM; TC2 — RAG or hybrid RAG+fine-tuned. Key contributions are: (1) a controlled paired evaluation of fine-tuning vs. RAG under identical model weights, revealing task-dependent adaptation regimes; (2) a multi-faceted automated evaluation framework with a six-item Directive 2023/970 checklist and RAGAS-inspired faithfulness metrics [12]; and (3) the empirical finding that TC1 and TC3 saturate at $n \approx 40$ training examples in total (≈ 13 per task)—a low-sample regime that establishes enterprise feasibility. Future work will address multilingual (Polish) deployment and evaluation on real HR documents.

ACKNOWLEDGMENT

During the preparation of this work, the authors used Claude (Anthropic) for language editing, phrasing suggestions and limited drafting assistance. GPT-4 (OpenAI) was used for synthetic training data generation (TC2, TC3 examples via constrained self-instruct, Section III-C). All outputs were reviewed and validated by the authors, who take full responsibility for the content of this publication.

REFERENCES

- [1] European Parliament and Council, "Directive (EU) 2023/970 of the European Parliament and of the Council on pay transparency and enforcement of the principle of equal pay," *Official Journal of the European Union*, L 132, pp. 21–44, 10 May 2023.
- [2] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., "Training Language Models to Follow Instructions with Human Feedback," in *Proc. 36th Conf. Neural Information Processing Systems (NeurIPS)*, 2022.
- [3] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai and Q. V. Le, "Finetuned Language Models Are Zero-Shot Learners," in *Proc. 10th Int. Conf. Learning Representations (ICLR)*, 2022.
- [4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., "Language Models are Few-Shot Learners," in *Proc. 34th Conf. Neural Information Processing Systems (NeurIPS)*, 2020, pp. 1877–1901.
- [5] S. G. Patil, T. Zhang, X. Wang and J. E. Gonzalez, "Gorilla: Large Language Model Connected with Massive APIs," *arXiv preprint arXiv:2305.15334*, 2023. DOI: 10.48550/arXiv.2305.15334.
- [6] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, B. Qian, S. Zhao, L. Hong, R. Tian, R. Xie, J. Zhou, M. Gerstein, D. Li, Z. Liu and M. Sun, "ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs," *arXiv preprint arXiv:2307.16789*, 2023. DOI: 10.48550/arXiv.2307.16789.
- [7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in *Proc. 10th Int. Conf. Learning Representations (ICLR)*, 2022.
- [8] T. Dettmers, A. Pagnoni, A. Holtzman and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," in *Proc. 37th Conf. Neural Information Processing Systems (NeurIPS)*, 2023.
- [9] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, M. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer and O. Levy, "LIMA: Less Is More for Alignment," in *Proc. 37th Conf. Neural Information Processing Systems (NeurIPS)*, 2023.
- [10] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo and Y. Iwasawa, "Large Language Models are Zero-Shot Reasoners," in *Proc. 36th Conf. Neural Information Processing Systems (NeurIPS)*, 2022.
- [11] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi and H. Hajishirzi, "Self-Instruct: Aligning Language Models with Self-Generated Instructions," in *Proc. 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023, pp. 13484–13508. DOI: 10.18653/v1/2023.acl-long.754.
- [12] S. Es, J. James, L. Espinosa-Anke and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," in *Proc. 18th Conf. European Chapter of the Association for Computational Linguistics (EACL)*, 2024, pp. 150–158. DOI: 10.18653/v1/2024.eacl-demo.16.
- [13] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," in *Proc. 2019 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410.
- [14] J. Johnson, M. Douze and H. Jégou, "Billion-Scale Similarity Search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, 2021. DOI: 10.1109/TBDATA.2019.2921572.
- [15] L. Xu, M. Skoularidou, A. Cuesta-Infante and K. Veeramachaneni, "Modeling Tabular Data using Conditional GAN" in *Proc. 33rd Conf. Neural Information Processing Systems (NeurIPS)*, 2019.
- [16] A. Vogelsang and M. Borg, "Requirements Engineering for Machine Learning: Perspectives from Data Scientists," in *Proc. 6th Int. Workshop on Artificial Intelligence for Requirements Engineering (AIRE)*, IEEE RE, 2019, pp. 245–251. DOI: 10.1109/REW.2019.00050.
- [17] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi and T. Zimmermann, "Software Engineering for Machine Learning: A Case Study," in *Proc. 41st IEEE/ACM Int. Conf. Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, 2019, pp. 291–300. DOI: 10.1109/ICSE-SEIP.2019.00042.
- [18] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, et al., "The Llama 3 Herd of Models," *arXiv preprint arXiv:2407.21783*, 2024. DOI: 10.48550/arXiv.2407.21783.
- [19] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand and G. Lengyel, "Mistral 7B," *arXiv preprint arXiv:2310.06825*, 2023. DOI: 10.48550/arXiv.2310.06825.
- [20] M. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, et al., "Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone," *arXiv preprint arXiv:2404.14219*, 2024.
- [21] S. N. Wood, *Generalized Additive Models: An Introduction with R*, 2nd ed. Boca Raton, FL: CRC Press, 2017.
- [22] I. Chalkidis, A. Jana, D. Hartung, M. Bommario, I. Androutsopoulos, D. M. Katz and N. Aletras, "LexGLUE: A Benchmark Dataset for Legal Language Understanding in English," in *Proc. 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022, pp. 4310–4330. DOI: 10.18653/v1/2022.acl-long.297.
- [23] I. Chalkidis, T. Pasini, S. Zhang, L. Tomada, S. F. Schwemer and A. Søgaard, "FairLex: A Multilingual Benchmark for Evaluating Fairness in Legal Text Processing," in *Proc. 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022, pp. 4389–4406. DOI: 10.18653/v1/2022.acl-long.301.