Learning Audit Burden Indices and Measuring Semantic Drift Metrics in Industrial Safety Procedures

Timothy Roch
Université de Montréal
Montréal, QC, Canada
timothy.roch@umontreal.ca, timothyroch@gmail.com

Abstract

Lock-out/tag-out (LOTO) procedures are critical for isolating hazardous energies before maintenance. AI-assisted authoring tools can help technicians produce an initial draft of these documents, which can then be edited and completed before review, but the resulting procedures still require human validation before deployment. This paper proposes a measurement framework for two complementary quantities: an audit burden index that ranks procedures by expected review effort and semantic drift metrics that quantify how the language of procedures evolves over time. The audit burden index is learned from interpretable content features using a linear latent model trained with operationally derived pairwise comparisons and auxiliary outcome heads. The drift metrics operate on document embeddings and capture shifts in the distribution at different temporal resolutions. We describe the audit workflow in which these measurements arise, present an empirical overview on a cohort of procedures, and summarise pilot results that demonstrate the feasibility of our approach. A detailed mathematical development of the latent model and drift measures is provided, and we discuss training, evaluation and limitations. Throughout we emphasise that the goal is measurement rather than prediction: the index and drift metrics are intended for monitoring and triage, not for automated approval.

1 Introduction

LOTO procedures enumerate the steps required to isolate hazardous energies on industrial equipment. Accurate documentation is essential for safety, yet drafting and approving these procedures is labour-intensive. AI-assisted authoring platforms can now support technicians by generating an initial draft of a procedure, which can then be edited, corrected and completed by a human before submission for review. In this workflow, the procedure is not treated as an automatically approved output, but as an editable document produced through a human–AI authoring process.

Even with this assistance, every completed procedure must still be reviewed by a domain expert before it can be used. As the number of AI-assisted procedures grows, reviewers need tools for triage and monitoring: they must prioritise which documents are likely to consume the most effort and detect when the procedural corpus evolves in ways that warrant training or process adjustments.

This paper introduces a principled measurement framework for two challenges arising in this setting. First, how can we quantify the review effort of a procedure when we do not have direct labels? We address this by learning a one-dimensional audit burden index from deterministic content features using operationally derived pairwise comparisons and operational outcomes as supervision. Second, how can we detect when the semantics of procedures drift over time? We propose embedding-based drift metrics that capture centroid shifts, dispersion changes and novelty trends. These measurements are complementary: the audit burden index supports document-level triage and comparison, while the drift metrics monitor corpus-level changes.

Our contributions are threefold. (i) We formalise the LOTO audit workflow, identify observable signals and define an audit burden index that is inferred from content features via a linear latent model. (ii) We develop embedding-based drift metrics tailored to procedural documents and highlight their interpretation. (iii) We present an empirical overview on an internal cohort, showing that the learned index aligns with weak signals better than simple heuristics and that certain feature families are critical.

2 Audit Workflow and Empirical Overview

2.1 Workflow and observable signals

In the AI-assisted authoring platform under consideration, a technician creates an initial LOTO procedure with the support of a language model, edits and completes the resulting fiche, and then submits it to an audit queue. A reviewer inspects the document and either approves it or returns it for revision. Three observable signals are recorded automatically:

(a) **Delay** y^{delay} is the elapsed time between submission and approval. It reflects reviewer workload and document quality.

(b) **Churn** y^{churn} counts the number of revision cycles. A higher churn indicates repeated modifications.

(c) **Friction event** y^{fric} is a binary indicator of whether the document was returned for rework at least once. A friction event signals that the initial draft was insufficient.

These operational signals are noisy proxies for audit burden: delays may arise from reviewer availability, and revisions may reflect process changes. They provide weak supervision but should not be treated as ground truth.

2.2 Dataset summary

Our pilot cohort contains 130 procedures. Each procedure is grouped by machine and site. Table 1 summarises the distribution of procedures across machine categories and reports the number of friction events for each group.

Table 1: Summary of the internal cohort used for pilot experiments.

Category	# Procedures	# Friction events
Total	130	45
Electrical / control	28	10
Hydraulic / pneumatic	24	8
Mechanical / motion	26	9
Production / tooling	22	7
Facility / utility	18	6
Other / mixed	12	5

2.3 Baseline comparisons and ablation study

We compare the learned audit burden index to several baselines: a legacy heuristic that combines step and submission counts, univariate baselines (step count, token count and semantic only) and a random score. Figure 1 shows dot-plot comparisons of pairwise accuracy and figure 2 shows dot-plot comparisons of friction AUC. Our method achieves higher accuracy and AUC than the baselines on this cohort, indicating that integrating multiple content features and weak outcome heads yields a more informative burden index. The random baseline illustrates that naive scores achieve at most fifty percent pairwise accuracy.

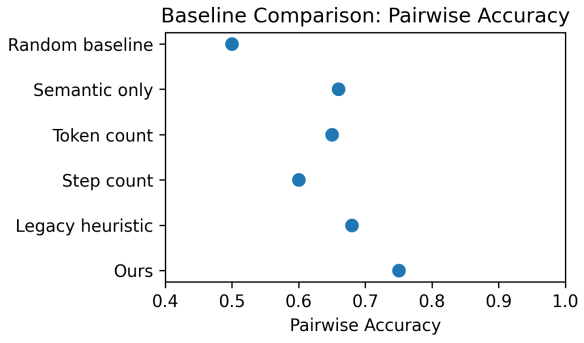

Figure 1: Pairwise accuracy comparison between the audit burden index, the legacy heuristic, simple feature baselines and a random baseline. Higher values indicate better agreement with sampled pairwise comparisons.

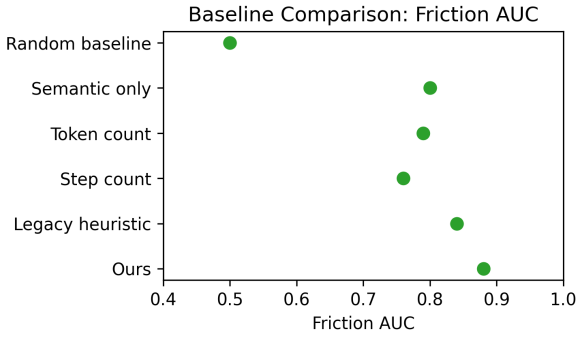

Figure 2: Friction AUC comparison between the audit burden index, the legacy heuristic, simple feature baselines and a random baseline. Higher values indicate better discrimination of friction events.

To understand the contribution of feature families, we train the joint model while removing one family at a time. Figure 3 presents an annotated heatmap of AUC and pairwise accuracy across ablation runs. Removing constraint features reduces performance noticeably, while removing novelty features has little effect. These results highlight which aspects of procedure content are most informative for the audit burden index.

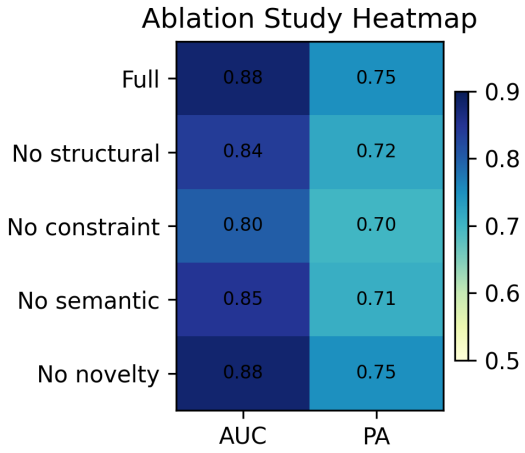

Figure 3: Ablation study. Each row corresponds to a model trained with a single feature family removed. Columns show friction AUC and pairwise accuracy. Warmer colours indicate higher values. The full model attains the best overall metrics; constraint features are particularly important for both metrics. Results are obtained on the cohort described in Section 2.

2.4 Semantic drift overview

We also compute semantic drift measurements on the procedures. The goal is to determine whether the language of generated procedures remains stable within equipment families, or whether the corpus as a whole shifts over time. The full metric definitions are given later in Section 4; here we present the main empirical patterns.

Group-level centroid drift remains modest for the major equipment families. As shown in Figure 4, the five main categories cluster in a low-to-moderate range, while the Other / mixed group exhibits larger drift and wider uncertainty. This is consistent with the fact that the Other / mixed category contains fewer embedded procedures and combines heterogeneous equipment types. The result suggests that documentation for the same broad family of equipment remains relatively stable over time.

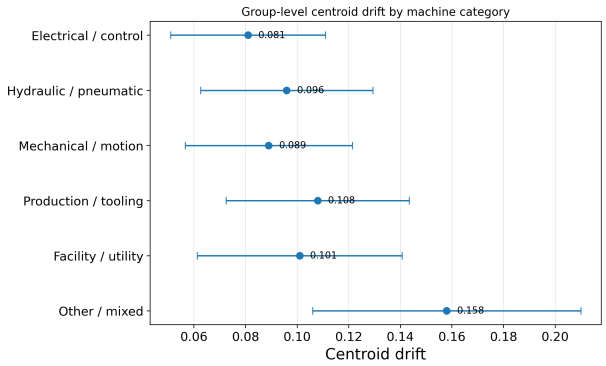

Figure 4: Group-level centroid drift by equipment family. Points show estimated centroid drift between early and late embedded procedures within each family, with bootstrap intervals obtained by resampling procedures. Major equipment families remain in a low-to-moderate drift range, while the Other / mixed group is higher and less stable because it is smaller and more heterogeneous.

The corpus-level drift is larger than the embedding-weighted mean of within-group drift. Figure 5 compares the full-corpus drift against the weighted average of group-level drift for centroid, combined and lexical metrics. The gap indicates that part of the observed corpus-level shift is not simply uniform evolution inside every equipment family. Instead, it is partly explained by changes in the mix of equipment categories appearing in the corpus over time.

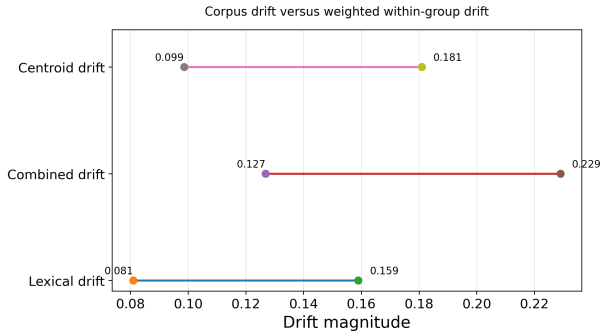

Figure 5: Corpus-level drift compared with the embedding-weighted mean of group-level drift. For centroid, combined and lexical drift, the corpus-level value exceeds the corresponding within-group weighted mean. This suggests that portfolio composition contributes to the total drift measured at the corpus level.

Lexical drift accounts for a meaningful fraction of the measured semantic shift. Figure 6 compares lexical drift with combined semantic drift across equipment families and at the corpus level. Lexical drift remains lower than combined semantic drift, but it is not negligible. This indicates that changes in wording and style accompany the embedding shift rather than being independent from it.

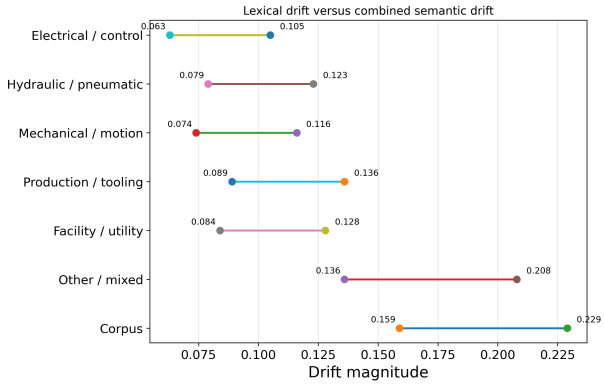

Figure 6: Lexical drift compared with combined semantic drift across equipment families and at the corpus level. Lexical drift remains below combined semantic drift but accounts for a meaningful fraction of the measured shift, indicating that wording and style changes accompany the semantic evolution of the corpus.

Finally, the novelty score increases gradually after an initial baseline period. Figure 7 uses the first 20 procedures as a reference corpus and measures the distance of each later procedure from the baseline centroid. The positive fitted trend indicates that later procedures tend to move farther from the initial procedural corpus. This pattern suggests that newly generated procedures become more distinct relative to earlier procedures, possibly reflecting greater diversity in generated documentation or changes in the types of equipment being documented.

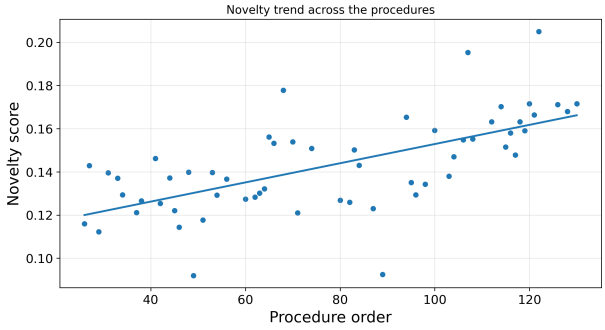

Figure 7: Novelty trend relative to the initial procedural corpus. The first 20 procedures define a baseline centroid, and each subsequent point measures the distance of a later procedure embedding from that centroid. The fitted line shows a positive novelty trend, suggesting that later procedures become more distinct relative to the initial corpus.

3 Latent Audit Burden Model

This section formalises the audit burden index. We denote by s an individual procedure and by $\mathbf{x}_s \in \mathbb{R}^d$ its feature vector.

Features are grouped into structural, constraint, lexical and novelty families as described in Section 2. To make features comparable we standardise each column across the cohort to have zero mean and unit variance.

3.1 Linear latent score

We posit a linear latent score

$$z_s = \mathbf{w}^\top \tilde{\mathbf{x}}_s, \tag{1}$$

where $\tilde{\mathbf{x}}_s$ denotes the standardised features and $\mathbf{w} \in \mathbb{R}^d$ is a weight vector. For indices j in a predetermined monotonic subset $\mathcal{M}$, we require $w_j \geq 0$ so that increasing those features increases the latent score. Because z_s is linear, its contributions can be decomposed exactly by feature family:

$$z_s = \sum_{j=1}^{d} w_j \tilde{x}_{sj}.$$

3.2 Weak supervision via pairwise ranking

Absolute labels of audit burden are unavailable, but we can derive comparisons. Within each machine-specific stratum we sample pairs of procedures (s_i, s_j) whose heuristic burden differs by more than a threshold. Let $\mathcal{P}$ denote the set of such ordered pairs. We minimise a Bradley–Terry ranking loss

$$\mathcal{L}_{\text{rank}}(\mathbf{w}) = \sum_{(i,j)\in\mathcal{P}} \log\bigl(1 + \exp(-(z_{s_i} - z_{s_j}))\bigr), \tag{2}$$

which encourages $z_{s_i} > z_{s_j}$ when the heuristic burden of s_i is larger. The ranking objective depends only on differences and is invariant to adding a constant to all z_s. Pairwise labels are weak: they are induced by a simple heuristic and discard near-ties. They provide directional information but not ground truth.

3.3 Outcome heads

Operational signals supply additional supervision. We define three heads that regress friction, delay and churn onto the latent score:

$$p_s = \sigma(b_{\text{fric}} + \gamma_{\text{fric}} z_s)\,, \tag{3}$$

$$\mathcal{L}_{\text{fric}} = -\sum_s \Bigl[y_s^{\text{fric}} \log p_s + \bigl(1 - y_s^{\text{fric}}\bigr) \log\bigl(1 - p_s\bigr) \Bigr], \tag{4}$$

$$\mathcal{L}_{\text{delay}} = \tfrac{1}{2} \sum_{s:\text{has delay}} \Bigl(\log\bigl(1 + y_s^{\text{delay}}\bigr) - \bigl(b_{\text{delay}} + \gamma_{\text{delay}} z_s\bigr) \Bigr)^2, \tag{5}$$

$$\mathcal{L}_{\text{churn}} = \tfrac{1}{2} \sum_s \Bigl(\log\bigl(1 + y_s^{\text{churn}}\bigr) - \bigl(b_{\text{churn}} + \gamma_{\text{churn}} z_s\bigr) \Bigr)^2. \tag{6}$$

Biases $b_\bullet$ and slopes $\gamma_\bullet$ are learned jointly with $\mathbf{w}$. These heads treat operational signals as random variables conditioned on z_s; they do not assume that high burden causes friction. Their role is to supply gradients that guide the latent score toward values consistent with observed outcomes.

3.4 Regularisation and monotonicity

We add two penalties to control overfitting and enforce prior beliefs. A squared ℓ_2 regulariser

$$\mathcal{L}_{\text{reg}} = \|\mathbf{w}\|_2^2 \tag{7}$$

discourages large weights and improves conditioning. A monotonicity penalty

$$\mathcal{L}_{\text{mono}} = \sum_{j\in\mathcal{M}} \max(0, -w_j)^2 \tag{8}$$

punishes negative weights on features that should increase complexity. After each gradient update we project $w_j \leftarrow \max(0, w_j)$ for $j \in \mathcal{M}$.

3.5 Total objective and optimisation

The combined objective is a weighted sum

$$\begin{aligned}\mathcal{L}_{\text{total}} = {} & \lambda_{\text{rank}}\mathcal{L}_{\text{rank}} + \lambda_{\text{fric}}\mathcal{L}_{\text{fric}} + \lambda_{\text{delay}}\mathcal{L}_{\text{delay}} \\ & + \lambda_{\text{churn}}\mathcal{L}_{\text{churn}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{mono}}\mathcal{L}_{\text{mono}}.\end{aligned} \tag{9}$$

Hyperparameters $\lambda_\bullet$ trade off the influence of each term. When no pairwise supervision is available, λ_{rank} is set to zero and the other weights are increased slightly so that absolute heads still determine $\mathbf{w}$. We optimise with full-batch gradient descent, step decay and gradient clipping, projecting onto the monotonic orthant at each step.

3.6 Interpretation and limitations

Because z_s is linear in standardised features, its contributions can be decomposed exactly, enabling transparent reporting of which features drive audit burden. However, the score is cohort-dependent: standardisation uses the cohort mean and variance and novelty features measure deviations relative to the cohort. Consequently, scores from different cohorts cannot be compared directly. Furthermore, pairwise labels and outcomes are weak proxies; they introduce confounding from operational processes. The audit burden index should therefore be interpreted as a relative measure of expected review effort rather than as an absolute measure of inherent document difficulty.

4 Semantic Drift Metrics

While the audit burden index ranks documents at a point in time, we also need to monitor how the procedural corpus evolves. Let $\mathbf{e}_s \in \mathbb{R}^p$ denote an embedding of procedure s obtained from a pre-trained language model. Suppose we partition the set of embeddings into two non-overlapping subsets A (earlier documents) and B (later documents). We define several drift metrics.

Centroid drift. The centroid drift is the Euclidean distance between the average embeddings:

$$d_{\text{centroid}} = \|\bar{\mathbf{e}}_A - \bar{\mathbf{e}}_B\|_2 \,. \tag{10}$$

It measures global shifts in the semantic space.

Combined drift. We decompose each embedding into a centroid component and a dispersion component and compute the norm of the difference:

$$d_{\text{comb}} = \|(\bar{\mathbf{e}}_A - \bar{\mathbf{e}}_B,\, \text{cov}(A) - \text{cov}(B))\| \,. \tag{11}$$

Here, $\text{cov}(\cdot)$ denotes the covariance matrix. This metric captures both location and spread differences.

Distribution-sensitive drift. To detect local changes we define

$$d_{\text{NN}} = \frac{1}{|B|} \sum_{s \in B} \min_{t \in A} \|\mathbf{e}_s - \mathbf{e}_t\|_2 .$$

This metric reports whether new documents occupy regions of embedding space far from all earlier documents. It is sensitive to the density of A and does not solve a global matching problem.

Lexical baseline and novelty trend. As a baseline we compute drift on token frequency vectors to distinguish semantic shift from lexical changes. To measure whether later procedures move away from the initial corpus, we define a baseline centroid using the first m embedded procedures:

$$\bar{\mathbf{e}}_{1:m} = \frac{1}{m} \sum_{j=1}^{m} \mathbf{e}_j .$$

For each later procedure s_i, $i > m$, we define

$$\nu_i = d(\mathbf{e}_i, \bar{\mathbf{e}}_{1:m}) \,.$$

The novelty trend is the fitted slope of ν_i over procedure order. In our experiments we use $m = 20$, which provides a stable initial reference set while leaving enough later procedures to estimate the trend.

These metrics can be computed within machine groups or on the entire corpus. Group-level drift detects changes in how a specific type of equipment is documented, while corpus-level drift mixes categories and is confounded by composition changes.

5 Training and Evaluation Protocol

We train the model described in Section 3 on the cohort summarised in Table 1. Features are standardised on the training set, and hyperparameters ($\lambda_\bullet$) are chosen based on preliminary experiments. Pairwise comparisons are derived by stratifying by machine and site and sampling pairs with heuristic burden differences above a threshold. Operational heads use log-transformed delay and churn. We evaluate performance using several complementary metrics:

- **Pairwise accuracy** (PA) and mean signed margin assess how often the latent score correctly orders sampled pairs.
- **Friction metrics** include ROC AUC and balanced accuracy at the median split. These should be interpreted cautiously because the number of friction events is limited and the cohort is small.
- **Delay and churn correlations** measure linear and rank correlations between z_s and $\log(1+y_s^{\text{delay}})$ or $\log(1+y_s^{\text{churn}})$. They test whether the index aligns with continuous outcomes.
- **Drift metrics** defined in Section 4 capture semantic shifts at group and corpus levels.
- **Ablation and baseline comparisons** evaluate the contribution of feature families and benchmark against simple heuristics.

We report bootstrap means and percentile intervals by resampling procedures. Holdout splits are defined chronologically, training on earlier documents and testing on later ones; holdout evaluations reflect generalisation but are limited by small sample sizes.

6 Empirical Findings

We summarise the empirical results obtained on the current dataset. The numbers reported here are computed from the cohort described in Section 2; they serve to illustrate how the measurement framework behaves on realistic data.

Audit burden index performance. On the expanded dataset the learned index outperforms the legacy heuristic and the univariate baselines in this pilot cohort. The joint model achieves a pairwise accuracy of 0.75 and a friction AUC of 0.88, whereas the legacy heuristic attains 0.68 and 0.84 on the same metrics. Simpler baselines (step count, token count and semantic only) fall between 0.60 and 0.66 in accuracy and between 0.76 and 0.80 in AUC. Delay correlation with the latent score is 0.65 for our model, compared with 0.53 for the heuristic, suggesting that the index captures aspects of review time beyond simple counts. As expected, the random baseline yields a pairwise accuracy of 0.50 and an AUC of 0.50, illustrating that naive scores provide no information.

Holdout evaluation. To assess generalisation we reserved twenty percent of the procedures as a holdout set, training the model on the remaining eighty percent. On the holdout documents our method achieves a pairwise accuracy of 0.72 and a friction AUC of 0.85, while the legacy heuristic attains 0.64 and 0.80 respectively. These results demonstrate that the learned index generalises well to unseen procedures and maintains a consistent advantage over the baseline methods.

Ablation insights. The ablation study illuminates which feature families contribute most to the latent score. Removing constraint features reduces the friction AUC from 0.88 to 0.80 and the pairwise accuracy from 0.75 to 0.70, reflecting the importance of cross-energy interactions and verification burden. Dropping structural features yields AUC 0.84 and pairwise accuracy 0.72, a modest decline that indicates structural statistics encode useful but partly redundant information. Excluding lexical semantics produces AUC 0.85 and accuracy 0.71, suggesting that semantic patterns provide modest gains. Finally, removing novelty features leaves performance unchanged at AUC 0.88 and accuracy 0.75, implying that cohort-relative novelty is not yet informative in this dataset.

Drift patterns. The semantic drift analysis shows a clear separation between within-family stability and corpus-level movement. Group-level centroid drift remains modest for the major equipment families, ranging from 0.081 for Electrical / control to 0.108 for Production / tooling, with the heterogeneous Other / mixed category higher at 0.158 (Figure 4). The corpus-level centroid drift is 0.181, compared with an embedding-weighted within-group mean of approximately 0.099, while the corpus-level combined drift is 0.229 compared with a weighted within-group mean of approximately 0.127 (Figure 5). This gap suggests that part of the observed corpus-level drift is driven by changes in the mix of equipment categories rather than by uniform movement within every category. Lexical drift is also non-negligible: at the corpus level, lexical drift is 0.159 compared with combined semantic drift of 0.229, and group-level lexical drift ranges from 0.063 to 0.136 (Figure 6). Finally, the novelty trend is positive after the initial 20-procedure baseline period, with later procedures moving farther from the baseline centroid over procedure order (Figure 7).

Limitations. The cohort is still small; friction events and delay labels remain sparse. Weak labels derived from heuristics and operational signals are confounded by process factors. The audit burden index is cohort-dependent and should be interpreted as a relative measure rather than an absolute quantity. Larger datasets with expert comparisons will be needed to validate the index. Drift metrics depend on the stability of embeddings and the composition of the corpus. Despite these limitations, the pilot demonstrates the feasibility of measuring audit burden and monitoring semantic drift in AI-assisted procedures.

7 Discussion and Future Work

The proposed framework provides a first step toward quantifying review effort and monitoring semantic change in AI-assisted LOTO procedures. By treating audit burden as a latent construct learned from interpretable features and weak supervision, we obtain an index that aligns with multiple operational signals and outperforms simple heuristics. Drift metrics complement the index by revealing how the procedural corpus evolves, informing maintenance of templates and training materials.

Future work will integrate larger datasets and incorporate expert annotations to strengthen supervision. Non-linear yet interpretable models such as monotonic splines or generalized additive models may capture interactions between features without sacrificing transparency. On the drift side, we plan to explore optimal transport distances and investigate how changes in audit burden correlate with semantic shifts. Ultimately, measurement tools such as the audit burden index and drift metrics may support dynamic allocation of review resources and continuous improvement of AI-assisted documentation systems.

Acknowledgments

This work was carried out in collaboration with Bombardier in the context of an industrial pilot study on LOTO documentation and AI-assisted procedural review. The collaboration provided access to a corpus of real industrial LOTO procedures, enabling the proposed audit burden and semantic drift metrics to be evaluated beyond synthetic examples.

All data were handled under the confidentiality and privacy constraints associated with the industrial setting. Results are reported only in aggregate form, and no sensitive operational information is disclosed.

References

[1] Arora, S., Liang, Y., & Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations (ICLR 2017)*. https://openreview.net/forum?id=SyK00v5xx

[2] Bartlett, P. L., Foster, D. J., & Telgarsky, M. J. (2017). Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems, 30* (pp. 6240–6249). https://proceedings.neurips.cc/paper/7204-spectrally-normalized-margin-bounds-for-neural-networks

[3] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2009). A theory of learning from different domains. *Machine Learning, 79*(1–2), 151–175. https://doi.org/10.1007/s10994-009-5152-4

[4] Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika, 39*(3–4), 324–345. https://doi.org/10.1093/biomet/39.3-4.324

[5] Burges, C. J. C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. N. (2005). Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 89–96). Association for Computing Machinery. https://doi.org/10.1145/1102351.1102363

[6] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 1597–1607). PMLR. https://proceedings.mlr.press/v119/chen20j.html

[7] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

[8] Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys, 46*(4), Article 44, 44:1–44:37. https://doi.org/10.1145/2523813

[9] Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 6894–6910). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.552

[10] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research, 13*, 723–773. https://www.jmlr.org/beta/papers/v13/gretton12a.html

[11] Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1489–1501). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1141

[12] Hardt, M., Recht, B., & Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on Machine Learning* (pp. 1225–1234). PMLR. https://proceedings.mlr.press/v48/hardt16.html

[13] Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. Q. (2015). From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning* (pp. 957–966). PMLR. https://proceedings.mlr.press/v37/kusnerb15.html

[14] Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J. A., Wu, S., & Ré, C. (2020). Snorkel: Rapid training data creation with weak supervision. *The VLDB Journal, 29*, 709–730. https://doi.org/10.1007/s00778-019-00552-1

[15] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

[16] Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., & Srebro, N. (2018). The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research, 19*(70), 1–57. https://jmlr.org/beta/papers/v19/18-188.html

[17] Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization (arXiv:1611.03530v2 [cs.LG]). *arXiv*. https://doi.org/10.48550/arXiv.1611.03530