DutyFormer: Neural-Symbolic Prediction of Silent Obligation Omissions in AI-Agent Process Traces

Mainak Basak
Hankyong National University
Anseong, Gyeonggi, Republic of Korea

Geon-Yun Shin*
gunyoon.shin@hknu.ac.kr
Hankyong National University
Anseong, Gyeonggi, Republic of Korea

Abstract

Modern Artificial Intelligent (AI) agentic systems execute multi-step process sequences encompassing tool utilization, observations, policy evaluations, and conclusive actions. A significant yet little examined failure mode is the *silent obligation omission*: the agent neglects to execute a necessary verification, approval, citation, escalation, or recording step prior to its deadline, without generating an explicit error signal. We provide DutyFormer, a neural-symbolic online monitor that forecasts omissions based on partial traces. The primary technique is *Symbolic-Attention-Bias* (SAB), which integrates a prefix-observable symbolic conformance vector directly into Transformer attention logits, enabling pending responsibilities, deadline slack, and conformance cost to influence sequence reasoning at each layer. We assess DutyFormer over four authentic process-log domains marked with timed obligation rules and externally validate the formulation using public benchmark-derived τ-bench tool-use trajectories. DF exhibits an obligation miss rate that rivals the most robust models, significantly surpassing all language-model baselines, and is distinguished by its unique calibration (conformal coverage), multi-task capabilities, and resilience to symbolic-state corruption. Consequently, we classify it as the most potent overall online supervision layer; yet, in this conformance-rich context, solely discriminative tabular models achieve comparable performance in terms of raw miss rate. Ablation studies indicate that symbolic attention refines the obligation choice border, whereas τ-bench validation demonstrates that the formulation corresponds to traces of tool-using agents. We examine the constraints of process-to-agent checkpoint transmission and the assessment of time-to-violation.

CCS Concepts

• **Computing methodologies** → **Neural networks**; • **Information systems** → *Process modeling languages.*

Keywords

process monitoring; obligation reasoning; neural-symbolic; transformer; conformance checking; AI-agent oversight

*corresponding author

KDD Workshop '26, Jeju, Korea

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

AI agents are increasingly deployed in high-stakes processes, loan adjudication, clinical triage, regulatory compliance, customer support, where they execute multi-step traces: they observe state, call tools, inspect outputs, decide whether to seek approval, and produce final actions. This setting exposes a qualitatively new failure mode that we call *silent obligation omission*: the agent selects a locally coherent action while silently failing to perform a procedurally mandated duty, for example, skipping verification after a tool output, omitting an approval before a side-effectful call, or failing to log evidence before a final response. Unlike overt errors, silent omissions produce no exception and no immediate complaint; they surface only when an audit or downstream incident occurs.

Existing methods do not directly target this problem. Predictive process monitoring [18] predicts trace outcomes or next events over unconstrained activity sequences, but does not model which obligations are triggered, how much deadline slack remains, or how severe a forthcoming breach would be. Conformance-checking methods [1, 16] detect violations *post hoc* and cannot produce forward-looking risk scores over future trace suffixes. What is missing is a monitor that turns prefix-observable obligation state into calibrated, online risk estimates usable inside an agent's execution loop. Given an observed process prefix $\tau^{(k)} = \langle e_1, \ldots, e_k \rangle$ and a domain obligation set, DutyFormer predicts whether a triggered obligation will be omitted by the end of the trace, that is, whether the obligation, once activated, will go unsatisfied past its deadline. Gold labels are computed deterministically by replaying the full trace through a formal obligation engine (DECLARE-style response constraints [15] compiled to Time Petri Nets [13]) and propagating the terminal Violated state back to each prefix. Crucially, only prefix-observable conformance state is exposed at runtime, so the model must predict from incomplete observations.

The principal contributions of our study are as follows.

(1) We formalize silent obligation omission detection as a structured online prediction problem grounded in deontic process constraints, with a prefix-level annotation protocol.
(2) We introduce SAB, which injects the current symbolic obligation state into attention logits rather than concatenating it as an input feature.
(3) On four process-log domains we show DutyFormer is a strong online neural-symbolic monitor. Its miss rate is competitive with the benchmark models in an order of magnitude below the language-model baselines we tested. It uniquely couples calibrated conformal coverage with multi-task outputs, and degrades gracefully under symbolic-state corruption, a combination no baseline matches.

(4) We externally validate the formulation on public benchmark-derived τ-bench [21] tool-use trajectories.

The rest of the paper is organized as follows: Section 1 introduces the problem. In Section 2 we outline the related works. Section 3 presents our proposed framework. The implementation details and experimental evaluations are discussed in Section 4, and Section 5 offers concluding remarks with discussion and limitations.

2 Related Work

Teinemaa et al. [18] provide a landmark survey and benchmark of outcome-oriented PPM, evaluating feature encodings and classifiers on eleven real-world logs. Sequence-to-sequence models for next-activity prediction were introduced by Evermann et al. [7] using LSTM, and extended by Theis et al. [19] to jointly predict next activity and remaining time. Camargo et al. [5] use generative models for full trace simulation. ProcessTransformer [3] applies a standard Transformer encoder to activity token sequences and achieves state-of-the-art next-event accuracy. DA-LSTM [19] introduces decay-augmented attention for handling irregular event timestamps. None of these works addresses obligation omission detection or integrates symbolic conformance knowledge into the model.

Classical conformance checking [16] computes an alignment cost between an observed trace and a reference Petri net. De Leoni et al. [6] extend alignment to data-aware models. Van der Aalst et al. [20] develop token-replay algorithms for conformance cost computation. DECLARE constraints [12, 15] form a declarative alternative, expressing process rules as LTL_f formulas over activity alphabets. These methods are post-hoc diagnostics; DutyFormer uniquely combines their formal structure with online predictive modelling.

Several recent studies fine-tune BERT-class models on activity sequences [9, 14] for compliance classification. Busch et al. [4] investigate prompt-engineering strategies for applying LLMs to business-process tasks, motivating our comparison against LLM-based process monitors. Neural discrete-time survival models [8, 10] frame time-to-event prediction as a sequence of Bernoulli hazard probabilities over time bins, enabling proper right-censored training.

Split conformal prediction [2, 17] provides finite-sample, distribution-free coverage guarantees. We apply it to wrap DutyFormer's omission probability with a calibrated prediction interval, enabling trustworthy threshold-free deployment.

Our configuration differs from standard predictive process monitoring because the monitor has access to a runtime conformance state computed from a formal obligation engine. This creates a high-information prediction regime: simple tabular models over conformance features can be highly competitive, and global discrimination metrics such as AUROC may saturate. We therefore evaluate DutyFormer not only by discrimination, but also by obligation-specific miss rates, calibration, ablation behavior, and deployability as an online oversight layer for AI-agent traces.

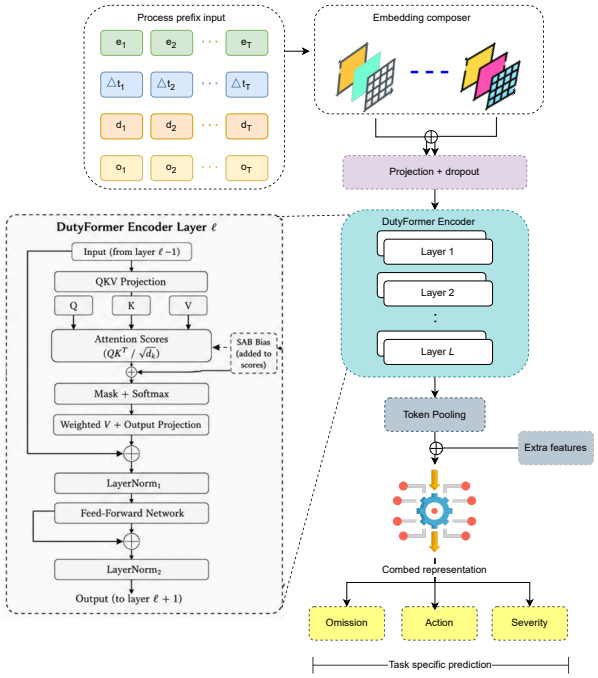


Figure 1: Overview of DutyFormer architecture.

3 Methodology

Figure 1 overviews DutyFormer. The model consumes a tokenized activity prefix together with continuous inter-event intervals, domain identity, per-token obligation-family assignments, and a symbolic conformance vector $\boldsymbol{s}_k$. It produces a multi-task output: an omission probability, a required-action class, and a violation-severity estimate.

After replaying a prefix $\tau^{(k)}$ through an ensemble of obligation Time Petri Nets, we extract a four-dimensional prefix-observable state $\boldsymbol{s}_k \in [0,1]^4$:

$$\boldsymbol{s}_k = \left(\mathbb{1}[\text{pending}>0],\ \mathbb{1}[\text{violated}>0],\ \frac{\delta_{\min}}{\Delta_{\max}},\ \min\!\left(1, \frac{c_k}{c_{\max}}\right) \right) \in [0,1]^4. \tag{1}$$

encoding whether any obligation is pending or already violated, the minimum remaining deadline slack $\delta_{\min}$, and the normalized cumulative conformance cost c_k.

The core novelty is that, rather than treating symbolic state as another input feature, SAB uses it to modulate the attention computation itself. In a standard head h, the scaled dot-product score for query i and key j is $\alpha_{ij}^{(h)} = \boldsymbol{q}_i^{(h)\top} \boldsymbol{k}_j^{(h)} / \sqrt{d_h}$. SAB augments every layer l with an additive symbolic bias derived from $\boldsymbol{s}_k$:

$$\alpha_{ij}^{(h,l)} = \frac{\mathbf{q}_i^{(h)\top} \mathbf{k}_j^{(h)}}{\sqrt{d_h}} + \beta^{(h,l)}\, m_j, \qquad \boldsymbol{\beta}^{(l)} = \tanh\!\left(W_{\text{sab}}^{(l)}\, \boldsymbol{s}_k\right), \tag{2}$$

where $W_{\text{sab}}^{(l)} \in \mathbb{R}^{H\times 4}$ is a layer-specific learnt projection producing per-head symbolic signal $\boldsymbol{\beta}^{(l)} \in [-1,+1]^H$, and $m_j \in \{0,1\}$ is a key-position obligation mask: $m_j = 1$ if key position j carries a live obligation-family tag $o_j \neq \varnothing$ and $m_j = 0$ otherwise. Equivalently,

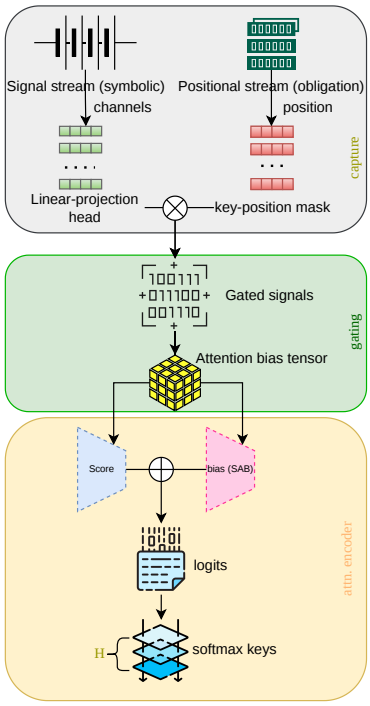


Figure 2: Overview of SAB module.

the SAB bias tensor $B^{(l)} \in \mathbb{R}^{H\times T\times T}$ of Figure 1 has entries $B_{ij}^{(h,l)} = \beta^{(h,l)} m_j$: the same per-head symbolic signal is applied to every query row but only into obligation-tagged key columns. The tanh bounds the bias to $[-1, +1]$, keeping the additive term within the numerical range of the attention logits.

Why a key-selective attention bias. A position-invariant scalar added to every (i, j) pair would be absorbed by the row-wise softmax and leave attention weights unchanged. Key-gating breaks this invariance: because the bias $\beta^{(h,l)} m_j$ varies across key positions j within each attention row, softmax redistributes mass *toward* (or away from) obligation-bearing keys depending on the sign of $\beta^{(h,l)}$. The symbolic state $\boldsymbol{s}_k$ thus gains layer-wise, per-head control over how strongly every query token attends to obligation-carrying keys, without inflating sequence length or adding a token. Alternative placements, concatenating $\boldsymbol{s}_k$ only at the pooled representation, or injecting it as an extra token gives the model access to conformance state but do not condition intermediate attention on obligation-bearing positions at every layer.

3.1 Multi-task output

The last non-padding token representation, concatenated with a small extra-feature vector of conformance scalars, feeds three heads: an omission head (sigmoid, trained with focal loss [11] for class imbalance), an action-family head (cross-entropy), and a severity regressor. The auxiliary heads act as a representation regularizer, encouraging the encoder to maintain an internal estimate of obligation state at short prefixes where the trigger has just fired. A split conformal wrapper [2] converts the omission probability into a distribution-free prediction set with finite-sample coverage, a guarantee none of our baselines provide. For training we also include a discrete-time survival head trained jointly with weight $\lambda_{\text{tte}} = 0.3$; in the current checkpoint this head collapses to a near-constant hazard and we make no time-to-violation claim (Appendix 10). The remaining three heads: omission, action, severity, are the deployed outputs.

Table 1: Evaluation data. Process logs are event logs; τ-bench is a public benchmark-derived agent-trajectory set used for external validation. Prefixes are generated at all lengths $k \in \{2, \ldots, n-1\}$.

Source	Trace type	Cases/Ep.	Prefixes	Omis.%
Process logs	real event logs	190,483	1,561,937	53.8
Incident Resp.	real	7,554	57,583	45.8
Loan Approval	real	31,509	1,079,537	68.4
Road Traffic	real	150,370	411,100	16.5
Sepsis	real	1,050	13,717	54.1
τ-bench	agent trajectories	1,980	70,877	83.4

4 Evaluation

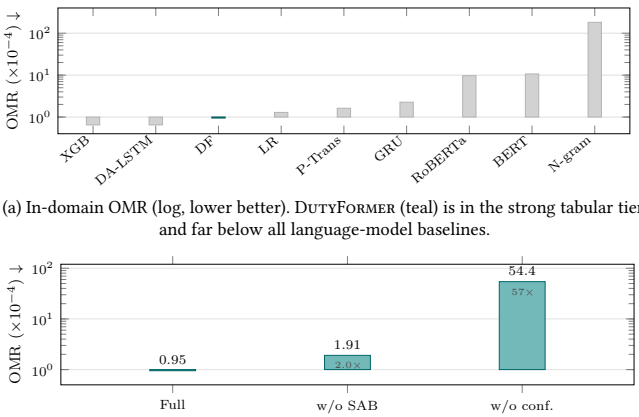
We evaluate on four real process-log domains from the BPI repositories, annotated by a domain-expert process analyst per domain with timed obligation families (escalation, approval, disclosure, halt/block) and severity weights. It is then prefix-labelled by Time-Petri-Net replay. Table 1 lists the four largest BPI-derived domains, which form the in-domain benchmark of Section 4 (Table 2, Group A). For the leave-one-domain-out transfer study (Table 7) the training corpus additionally includes three further BPI-derived domains, healthcare billing, permit workflow, and procurement, each of which is held out in turn, giving the seven held-out domains reported there. Annotations were performed by a single analyst per domain following, a written obligation-authoring protocol. Formal inter-annotator agreement studies and multi-annotator label reconciliation are considered as future work and constitute a known scope limitation of the current benchmark. To test transfer to tool-using agents, we externally validate on public benchmark-derivedτ-bench [21] trajectories (airline and retail), normalized into the same prefix-monitoring schema (Table 1). These are benchmark-derived agent trajectories, not production enterprise logs. Obligation specifications by multi-annotator agreement studies and obligation-set expansion are our future work.

Metrics. Because the monitor has runtime access to a formal conformance state, this is a high-information regime in which global discrimination (AUROC) saturates (≥ 0.999 for all conformance-aware models). We therefore foreground *obligation-specific* metrics: the Obligation Miss Rate (OMR, fraction of truly obligated cases not flagged), its severity-weighted variant (SW-OMR), the False

Table 2: Compact evaluation, grouped by *purpose*. Group A: conformance-rich benchmark; Group B: SAB ablation (DutyFormer's primary evidence); Group C: τ-bench external validation. ↓ lower better, ↑ higher better.

Model	OMR↓	FIR↓	AUPRC↑	ECE↓	Reading
A. Full conformance benchmark (process logs)					
XGBoost	0.65	0.0012	1.000	0.0003	strong tabular base
DA-LSTM	0.65	0.0012	1.000	0.0006	strong sequence base
DutyFormer	0.95	0.0028	1.000	0.0017†	neural-symbolic monitor
RoBERTa-PM	9.70	0.0035	1.000	0.0015	language-model base
N-gram	182.0	0.0349	0.993	0.0261	weak sequence base
B. Ablation (DutyFormer)					
DutyFormer	0.95	0.0028	1.000	0.0017†	full model
w/o SAB	1.91	0.0030	1.000	–	tests symbolic attn.
w/o conf. feat.	54.4	0.281	0.995	0.132	tests symbolic state
C. τ-bench external validation (agent traces)					
DutyFormer	0.07*	0.024	0.999	0.009	agent-trace validation
w/o SAB	0.08*	0.018	1.000	0.008	ablation on agent traces
XGBoost	0.00	0.000	1.000	0.000	conformance-rich bound

†after temperature scaling (T=0.25; 11× ECE reduction). *τ-bench OMR in native units ($\times 10^{-1}$): 0.07=0.007.



(a) In-domain OMR (log, lower better). DutyFormer (teal) is in the strong tabular tier and far below all language-model baselines.

(b) Removing SAB doubles OMR and removing the conformance vector raises it 57× relative to the full model.

Figure 3: Result summary on the four-domain process-log benchmark.

Intervention Rate (FIR), AUPRC, and Expected Calibration Error (ECE).

We observe that since labels are computed by replaying the *full* trace through the obligation engine while s_k observes only the prefix, a natural concern is whether the task reduces to reading off an oracle already present in the input, and we argue that it does not. The label asks whether an obligation active at prefix k will *still* be unsatisfied at trace end; s_k records only what is currently pending. Most pending obligations are ultimately satisfied before their deadline. Therefore, among prefixes with an active pending obligation, both violated-outcome and satisfied-outcome cases occur with non-trivial frequency, and the model must forecast which resolution the remaining trace suffix will produce using activity vocabulary, deadline slack, and behavioural pattern. Two controls in Appendix 6 reinforce this: (i) an activity-history-only probe (no symbolic input) reaches AUROC 0.937, which is well above chance yet well below symbolic-aware models. Hence neither signal is trivially sufficient alone; and (ii) shuffled-label training collapses to AUROC 0.305, excluding label leakage through the feature pipeline.

Main results. Table 2 groups the evaluation by *purpose*. In Group A the full-conformance configuration is not a pure neural leaderboard: all conformance-aware models receive prefix-observable replay states, so purely discriminative tabular models such as XGBoost are expected to be highly competitive on raw miss rate. We read this as confirmation that obligation replay is a high-information signal and that the task is conformance-rich. Within this regime DutyFormer is the strongest *neural* monitor and beats every language-model baseline by an order of magnitude (OMR 0.95 vs. 9.70-182$\times 10^{-4}$; Figure 3). Group B isolates the mechanism: removing SAB *doubles* OMR with AUROC unchanged, and removing the symbolic conformance vector raises OMR 57× (Figure 3), showing symbolic attention and explicit conformance state are complementary. Group C shows the same formulation transfers to tool-using agent trajectories.

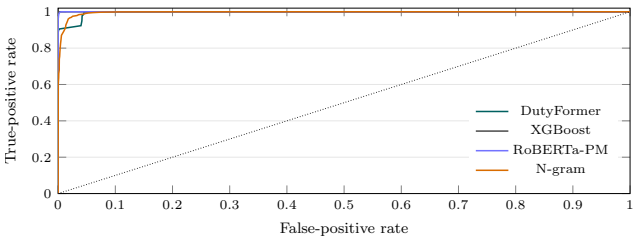
Although headline AUROC saturates (≥ 0.999 for all conformance-aware models), the underlying ROC and precision-recall geometry is *not* identical across models (Figure 4). XGBoost and the language-model baselines reach near-maximal true-positive rate at essentially zero false-positive rate. DutyFormer exhibits a distinct, informative shape: it captures roughly seven-eighths of violation-destined prefixes (TPR $\approx$ 0.88) at zero false alarms, then requires a small false-positive budget to recover the long tail. This is the empirical signature of the conformance-rich regime, the symbolic state cleanly separates the bulk of cases, leaving a hard residual that no model resolves without trading specificity and it motivates evaluating monitors by operating-point behaviour rather than by a single saturated scalar.

The more stringent setting is pre-violation early warning, where the model is evaluated only on prefixes in which the eventual omission is not yet directly observable in the input. This is the operationally meaningful regime for an online monitor, and DutyFormer retains positive discrimination (AUROC $\approx$0.72 across horizons of 1-3 events before any visible violation marker; Table 5 in Appendix 7). Sample sizes at short horizons are modest ($n \approx$ 70-80. Expanding this evaluation with a larger held-out pool of pre-violation prefixes is our follow-up study in future direction.

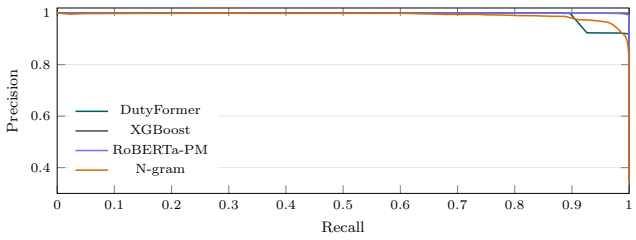
Operating-point behaviour. Figure 5 sweeps the decision threshold τ for DutyFormer and plots the resulting miss rate (OMR) and false-intervention rate (FIR). The two rates are cleanly separable: across a wide band of low-to-moderate thresholds the monitor holds OMR at or near zero while FIR falls below 5%, and the conformal wrapper picks an operating point inside this band automatically. This separability, rather than a marginal AUROC difference, is what makes DutyFormer usable as an online oversight layer with a tunable risk posture and finite-sample conformal coverage.

4.1 Robustness, Calibration, and Capability

Raw miss rate alone understates what an oversight layer must deliver. A deployable monitor must remain calibrated, hold coverage guarantees, and degrade gracefully when its symbolic inputs are imperfect. We evaluate our model on these axes and confirm



(a) ROC curves on test-set data.



(b) Precision-recall curve on test sets.

Figure 4: Discrimination geometry on the in-domain test split.

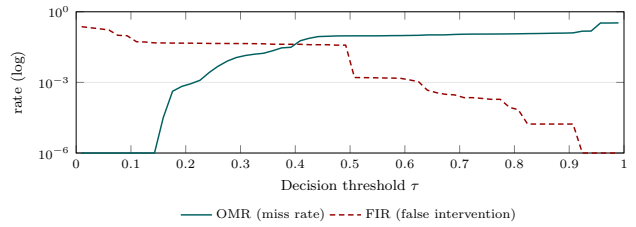


Figure 5: DutyFormer operating-point data. A wide band of thresholds holds the miss rate near zero while the false-rate stays low; the conformal method chooses an operating point inside this band.

that our model is the only framework in evaluation where all of these properties are measured and confirmed, with comparative robustness measurement left to future work.

We corrupt the prefix-observable symbolic inputs (masking obligation state, flipping flags, dropping deadline slack, swapping families) at increasing rates and measure operational degradation. DutyFormer degrades gracefully: as corruption rises from 0% to 20%, its ECE rises only from 0.009 to 0.026 while split-conformal coverage stays near the nominal 90% (0.990 → 0.974), so alerts remain trustworthy under noisy conformance extraction. Robustness is measured on the agent-trace validation set, where corrupted-input runs were available; extending the corruption protocol to all process-log baselines is left to future work.

Table 3 compares the operational capabilities each model provides. DutyFormer is the only method that combines online prefix prediction, symbolic-state fusion inside attention, multi-task action/severity outputs, temperature-scaled calibration, a distribution-free conformal-coverage guarantee, and measured robustness to symbolic noise.

Table 3: Operational capability comparison. Yes (supported by the model directly), Partial (supported via an external wrapper only), No (not supported), N/M (not measured). "N/M" cells reflect the scope of the present study, not architectural absence.

Capability	XGB/LR	GRU/P-Trans	BERT/RoBERTa	DutyFormer
Online prefix prediction	Yes	Yes	Yes	Yes
Symbolic conformance input	Yes	Yes	Partial	Yes
Symbolic state *in attention* (SAB)	No	No	No	Yes
Multi-task (action + severity)	No	Partial	No	Yes
Temperature-scaled calibration	Partial	Partial	Partial	Yes
Conformal coverage guarantee	No	No	No	Yes
Graceful symbolic-noise robustness	N/M	N/M	N/M	Yes

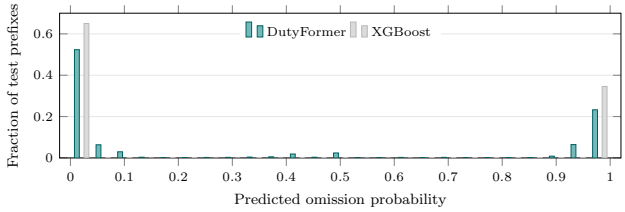


Figure 6: Predicted-score distribution on the held-out test split.

DutyFormer's raw probabilities are sharply bimodal, a direct consequence of the high mutual information between conformance state and label, with the bulk of mass at the extremes and a thin band of genuinely uncertain prefixes in between (Figures 6-7). Uncalibrated ECE is 0.0193; a single temperature parameter (T=0.25) fitted on validation reduces this 11× to 0.0017 and the Brier score to 0.0016, after which DutyFormer is competitive with the best-calibrated tabular baseline. The split-conformal wrapper then converts the calibrated score into a prediction set with finite-sample 90% coverage.

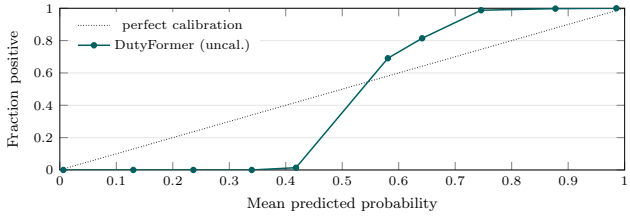


Figure 7: DutyFormer reliability diagram on the held-out test split.

4.2 Generalization

We assess how each model generalizes from its training domains to unseen ones via leave-one-domain-out (LODO) transfer, the closest analog to a held-out generalization test (full per-domain results in Appendix 8). Figure 8 contrasts in-domain test performance with cross-domain transfer, measured by Task Success under Compliance (TSuC, the fraction of cases that are simultaneously

task-successful and non-violating). We report TSuC rather than AUROC because, as discussed in Section 4, AUROC saturates in this conformance-rich regime, so an obligation-specific operational metric is more informative. In-domain, all three models are essentially tied at the ceiling (TSuC ≥ 0.998). DutyFormer matches the strongest tabular baselines and outperforms all language-model baselines. In cross-domain, XGBoost transfers with negligible loss (TSuC 0.9974) because it operates purely on vocabulary-agnostic conformance features; LR drops modestly to 0.9569; DutyFormer drops to 0.7791($\Delta \approx 0.22$) because its fixed activity-embedding table maps unseen tokens to an out-of-vocabulary embedding, injecting structured noise into every prefix representation. The gap is a property of the current token encoder, not of SAB, which halves in-domain OMR (Table 2) and operates on the vocabulary-agnostic conformance path. Vocabulary-free event encoders are the natural fix and we leave them to future work.

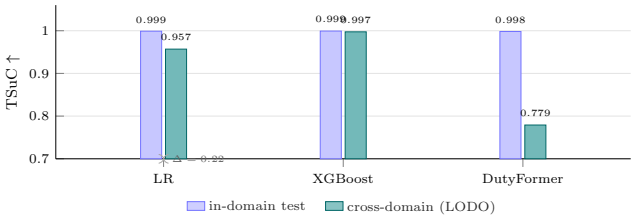


Figure 8: Generalization on Task Success under Compliance (TSuC): in-domain test vs. leave-one-domain-out (LODO) macro-average across seven held-out domains.

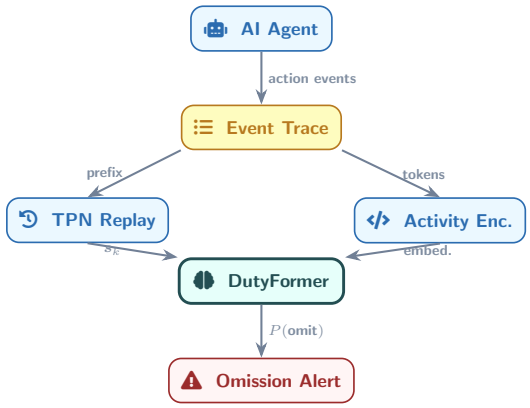


Figure 9: Process of DutyFormer module as an online compliance monitor inside an AI-agent pipeline.

4.3 Attention-Shift Diagnostic

The SAB mechanism is designed to redistribute attention toward obligation-bearing key positions. We verify this directly by comparing per-layer attention distributions between SAB-active and SAB-ablated forward passes on identical inputs. On a held-out sample of $N = 300$ prefixes drawn evenly across the four process-log domains, we capture post-softmax attention $A^{(h,l)} \in \mathbb{R}^{T \times T}$ at every encoder layer under both conditions and compute the difference map $\Delta A_{ij}^{(h,l)} = A_{ij}^{(h,l)}\big|_{\text{SAB}} - A_{ij}^{(h,l)}\big|_{\text{ablated}}$.

Figure 10 illustrates a representative layer/head pair from a loan-approval trace with $s_k = [1, 1, 0.0, 1]$ (pending, violated, zero deadline slack, non-trivial cost). In (a), the SAB-ablated attention shows a broadly diagonal pattern with weak mass on obligation keys. In (b), with SAB active and $\beta^{(3,1)} = +1.00$ under key-gating, attention mass is pulled off the diagonal and onto violated-obligation key positions (red vertical guides at $j \in \{18, 22, 24, 26, 28\}$). The difference map (c) makes the redistribution explicit. Red cells indicate keys receiving more mass under SAB, blue cells indicate keys ceding mass. The redistribution is confined to obligation-tagged columns exactly as the mechanism predicts; untagged columns are essentially unchanged ($|\Delta A| < 0.05$ almost everywhere off the columns).

5 Conclusion

In this study, the evidence supports a bounded claim. In the conformance-rich regime, DutyFormer matches strong tabular and sequence baselines (XGBoost, DA-LSTM) on raw obligation miss rate and lies an order of magnitude below the language-model baselines we tested. On top of that shared discriminative signal, DutyFormer supplies the per-alert operating machinery an online oversight layer needs, temperature-scaled calibration, split-conformal coverage, action-family routing, and severity triage, as first-class outputs of a single model rather than as external wrappers assembled around a discriminator. The SAB ablation (Table 2, Group B) isolates the position-selective attention bias as causally responsible for the miss-rate change. We do not claim uniqueness on any individual wrapper technique, temperature scaling, split conformal prediction, and multi-task outputs can in principle be attached to strong tabular baselines. We claim that this consolidation into one calibrated, coverage-guaranteed model is what the deployment setting demands, and that we are the first to demonstrate it end-to-end for silent obligation omission.

Several limitations remain in our study. First, direct process-to-agent checkpoint transfer is limited by vocabulary shift: a process-trained DutyFormer applied zero-shot to τ-bench activities degrades because its fixed activity-embedding table maps unseen tokens to an out-of-vocabulary embedding, whereas the vocabulary-agnostic conformance features transfer well (Appendix 9). Second, τ-bench is benchmark-derived, not production; results are external validation of the formulation, not deployment evidence on live LangGraph, browser-agent, or enterprise RPA logs. Third, time-to-violation estimation remains exploratory: a discrete-time survival head exhibits constant-hazard collapse under the current multi-task weighting, so we make no survival claim and report the diagnostic only in Appendix 10. Finally, our obligations are expressible as DECLARE-response constraints; richer temporal modalities (inter-obligation precedence, data-conditional exemptions, nested deadlines) are not yet represented.

Future work. Our highest-priority next step is closing the cross-domain gap identified in Figure 8 by replacing the fixed activity-embedding table with a vocabulary-free encoder (byte- or subword-level tokenization, or conformance-family-derived embeddings). Because the conformance path already transfers near-perfectly

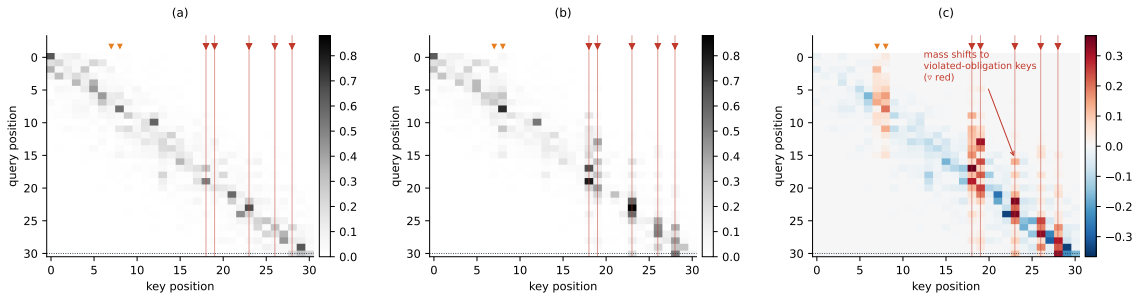


Figure 10: SAB attention-shift. (a) w/o SAB (b) with SAB active (c) Difference map

(Appendix 8), we expect this change to substantially recover cross-domain TSuC without altering the SAB mechanism itself. Beyond that, we plan (i) multi-baseline symbolic-noise robustness sweeps to substantiate the comparative robustness column in Table 3, (ii) multi-seed replication with bootstrap CIs on the SAB ablation, and (iii) evaluation on live agent traces, browser-agent systems, and enterprise RPA logs with human-verified obligation failures. We view silent obligation omission as a distinct, useful online-monitoring target for trustworthy AI-agent deployment and SAB as a lightweight, position-selective mechanism for injecting symbolic obligation state into sequence reasoning.

Acknowledgments

This work was supported by the Digital Columbus Project through the Institute of Information & Communications Technology Planning & Evaluation (IITP) funded by the Ministry of Science and ICT (RS-2026-25526071), and was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-25122978).

References

[1] Arya Adriansyah, Boudewijn F. van Dongen, and Wil M. P. van der Aalst. 2011. Towards Robust Conformance Checking. In *Business Process Management Workshops (Lecture Notes in Business Information Processing, Vol. 66)*. Springer, Berlin, Heidelberg, 122–133. doi:10.1007/978-3-642-20511-8_11

[2] Anastasios N. Angelopoulos and Stephen Bates. 2023. Conformal Prediction: A Gentle Introduction. *Foundations and Trends in Machine Learning* 16, 4 (2023), 494–591. doi:10.1561/2200000101

[3] Zaharah Allah Bukhsh, Aaqib Saeed, and Remco M. Dijkman. 2021. ProcessTransformer: Predictive Business Process Monitoring with Transformer Network. *ArXiv* abs/2104.00721 (2021). https://api.semanticscholar.org/CorpusID:233004463

[4] Kiran Busch, Alexander Rochlitzer, Diana Sola, and Henrik Leopold. 2023. Just Tell Me: Prompt Engineering in Business Process Management. In *BPMDS/EMMSAD@CAiSE*. https://api.semanticscholar.org/CorpusID:258170067

[5] Manuel Camargo, Marlon Dumas, and Oscar González-Rojas. 2019. Learning Accurate LSTM Models of Business Processes. In *Proceedings of the 17th International Conference on Business Process Management (Lecture Notes in Computer Science, Vol. 11675)*. Springer, Cham, Switzerland, 286–302. doi:10.1007/978-3-030-26619-6_19

[6] Massimiliano de Leoni and Wil M. P. van der Aalst. 2013. Aligning Event Logs and Process Models for Multi-Perspective Conformance Checking: An Approach Based on Integer Linear Programming. In *Business Process Management (Lecture Notes in Computer Science, Vol. 8094)*. Springer, Berlin, Heidelberg, 113–129. doi:10.1007/978-3-642-40176-3_10

[7] Joerg Evermann, Jana-Rebecca Rehse, and Peter Fettke. 2017. Predicting Process Behaviour Using Deep Learning. *Decision Support Systems* 100 (2017), 129–140. doi:10.1016/j.dss.2017.04.003

[8] Michael F. Gensheimer and Balasubramanian Narasimhan. 2019. A Scalable Discrete-Time Survival Model for Neural Networks. *PeerJ* 7 (2019), e6257. doi:10.7717/peerj.6257

[9] Jonghyeon Ko and Marco Comuzzi. 2023. A Systematic Review of Anomaly Detection for Business Process Event Logs. *Business & Information Systems Engineering* 65, 4 (2023), 441–462. doi:10.1007/s12599-023-00794-y

[10] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. 2019. Time-to-Event Prediction with Neural Networks and Cox Regression. *Journal of Machine Learning Research* 20, 129 (2019), 1–30.

[11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2999–3007.

[12] Fabrizio Maria Maggi, Marco Montali, Michael Westergaard, and Wil M. P. van der Aalst. 2011. Monitoring Business Constraints with Linear Temporal Logic: An Approach Based on Colored Automata. In *Business Process Management (Lecture Notes in Computer Science, Vol. 6896)*. Springer, Berlin, Heidelberg, 132–147. doi:10.1007/978-3-642-23059-2_13

[13] Philip Meir Merlin. 1974. *A Study of the Recoverability of Computing Systems*. PhD thesis. University of California, Irvine.

[14] Vincenzo Pasquadibisceglie, Annalisa Appice, Giovanna Castellano, and Donato Malerba. 2019. Using Convolutional Neural Networks for Predictive Process Analytics. In *Proceedings of the 1st International Conference on Process Mining*. IEEE, 129–136.

[15] Maja Pešić, Helen Schonenberg, and Wil M. P. van der Aalst. 2007. DECLARE: Full Support for Loosely-Structured Processes. In *Proceedings of the 11th IEEE International Enterprise Distributed Object Computing Conference*. IEEE, 287–300. doi:10.1109/EDOC.2007.14

[16] Anne Rozinat and Wil M. P. van der Aalst. 2008. Conformance Checking of Processes Based on Monitoring Real Behavior. *Information Systems* 33, 1 (2008), 64–95. doi:10.1016/j.is.2007.07.001

[17] Glenn Shafer and Vladimir Vovk. 2008. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research* 9 (2008), 371–421.

[18] Irene Teinemaa, Marlon Dumas, Marcello La Rosa, and Fabrizio Maria Maggi. 2019. Outcome-Oriented Predictive Process Monitoring: Review and Benchmark. *ACM Transactions on Knowledge Discovery from Data* 13, 2, Article 17 (2019), 57 pages. doi:10.1145/3301300

[19] Julian Theis and Houshang Darabi. 2019. Decay Replay Mining to Predict Next Process Events. *IEEE Access* 7 (2019), 119787–119803. doi:10.1109/ACCESS.2019.2937085

[20] Wil M. P. van der Aalst. 2016. *Process Mining: Data Science in Action* (2 ed.). Springer, Berlin, Heidelberg. doi:10.1007/978-3-662-49851-4

[21] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. *arXiv preprint arXiv:2406.12045* (2024). arXiv:2406.12045 [cs.AI] doi:10.48550/arXiv.2406.12045

Appendix: Ablation Results

The following ablation studies support the main claims but are not required to follow the argument. They are retained for transparency and reproducibility.

6 Anti-Leakage Evaluation

Because symbolic conformance features may appear close to the target definition, we run feature-removal and randomization probes (Table 4). All-feature and symbolic-only probes are near-perfect;

crucially, the shuffled-label probe collapses toward chance (AUROC 0.305), and the activity-history-only probe is far weaker than the symbolic-aware ones, indicating the signal is genuine prefix-observable conformance state rather than a trivial outcome leak.

Table 4: Anti-leakage audit (logistic probe). Shuffled-label performance should collapse toward chance, and it does.

Probe	AUROC	OMR	ECE
All features	1.000	0.000	0.004
Remove outcome columns	1.000	0.000	0.008
Remove future/post-event columns	0.979	0.000	0.069
Activity-history only (no symbolic)	0.937	0.429	0.112
Symbolic conformance only	1.000	0.000	0.024
Shuffled labels (sanity)	0.305	0.857	0.317
Shuffled symbolic states	0.913	0.429	0.094

7 Pre-Violation Early Warning

The pre-violation protocol evaluates only prefixes in which the eventual omission is not yet directly observable, distinguishing predictive monitoring from post-hoc conformance checking. DutyFormer retains positive discrimination (AUROC $\approx$ 0.72) even when evaluated several events *before* any visible violation marker (Table 5), evidence that it learns early temporal-symbolic risk patterns rather than merely echoing an already-visible violation.

Table 5: Pre-violation horizon evaluation (DutyFormer-compatible probe).

Horizon	N	AUROC	AUPRC	F1	OMR
1 event before	79	0.721	0.473	0.725	0.138
2 events before	74	0.724	0.436	0.700	0.125
3 events before	69	0.726	0.391	0.692	0.053
first 25% before	71	0.729	0.412	0.714	0.048
first 50% before	63	0.722	0.320	0.600	0.077

8 Cross-Domain Transfer (LODO)

Table 6 reports leave-one-domain-out transfer macro-averaged over seven usable held-out domains. Conformance-only classifiers transfer near-perfectly; DutyFormer trails owing to fixed-vocabulary embedding noise on unseen activities, as discussed in Section 4.2.

Table 6: Cross-domain transfer (leave-one-domain-out, macro-avg over 7 held-out domains). ↓ OMR/SW-OMR, ↑ AUROC/TSuC.

Model	AUROC	OMR	SW-OMR	TSuC
Logistic Regr.	0.9976	0.0558	0.0555	0.9569
XGBoost	0.9997	0.000	0.000	0.9974
DutyFormer	0.9624	0.0588	0.0587	0.7791

Table 7 gives the per-domain breakdown for DutyFormer. Transfer is strongly domain-dependent: it is near-perfect where the held-out activity vocabulary overlaps the training union (healthcare billing, loan approval) and degrades where out-of-vocabulary rates are high (road traffic AUROC 0.806; elevated FIR on incident response and permit workflow). This pattern is consistent with the vocabulary-shift diagnosis rather than a failure of the symbolic mechanism, whose conformance path is vocabulary-agnostic.

Table 7: Per-domain leave-one-domain-out transfer for DutyFormer data.

Held-out domain	n_{test}	AUROC	F1	OMR	FIR
Healthcare billing	7,056	0.9972	0.9516	0.000	0.042
Incident response	10,000	0.9849	0.7164	0.000	0.792
Loan approval	10,000	0.9961	0.9048	0.013	0.195
Permit workflow	10,000	0.9917	0.7002	0.000	0.856
Procurement	10,000	0.9897	0.8212	0.297	0.010
Road traffic	10,000	0.8056	0.7222	0.000	0.769
Sepsis workflow	10,000	0.9716	0.9460	0.102	0.000

9 Process-to-Agent Transfer

Applying a process-trained DutyFormer checkpoint zero-shot to τ-bench activities maps unknown tokens to an out-of-vocabulary embedding and degrades sharply (AUROC 0.453), whereas a feature-space logistic probe on the same vocabulary-agnostic conformance features reaches AUROC 0.999. We interpret this as a stress test that makes the vocabulary-shift limitation explicit, not as evidence against the formulation; agent-only training recovers strong performance (Table 2, Group C).

10 Survival Head Analysis

The discrete-time survival head is designed to produce per-prefix hazard curves. In the current checkpoint it collapses to a near-constant hazard function insensitive to the omission score, consistent with loss-weight imbalance (λ_{om}=1.0 vs. λ_{tte}=0.3) and right-censoring of $\sim$ 50% of prefixes. A survival-dedicated variant selected by validation still attains only C-index 0.295 and survival NLL 0.251. We therefore make *no* time-to-violation claim and treat survival as future work requiring dedicated rebalancing, monotonicity constraints, or a stratified survival loss. The scalar omission probability remains the appropriate decision variable in our study.