Explaining Concept Drifts in Business Processes: An LLM-Based Approach to Context-Aware Sense-Making in Process Mining

Jan Schaffner[1], Sandro Franzoi[1], and Jan vom Brocke[1]

University of Münster, Münster, Germany
{jan.schaffner, sandro.franzoi, jan.vom.brocke}@uni-muenster.de

Abstract. Concept drift detection in process mining identifies process changes but often fails to explain their underlying causes, creating a detection-interpretation gap. We addressed this problem by designing and evaluating the Concept Drift Explainer, a prototypical artifact that provides context-aware explanations for drifts. Using a Design Science Research methodology, we developed an LLM-based multi-agent system that systematically links drift signals from event logs to evidence from unstructured enterprise knowledge sources. We evaluated the artifact through a controlled experiment using real-world event logs and a synthetic context corpus. It achieved high retrieval accuracy, with Recall@2 = 100% and MRR = 0.875 across all evaluated drifts. Subsequent expert evaluation confirmed the artifact's practical utility, demonstrating its ability to substantially enhance analytical efficiency and explainability. In doing so, this paper contributes to research by operationalizing context-aware sense-making in process mining.

Keywords: Process Mining · Concept Drift · Large Language Models · Context · Explanation

1 Introduction

Process mining has evolved into a key technique for analyzing business processes based on digital trace data from information systems [3]. While the field has produced major advances in process discovery, conformance checking, and enhancement [3], it still struggles to provide meaningful explanations for *why* process changes occur [23]. This gap becomes especially pronounced in the context of concept drifts, where processes change systematically over time due to evolving organizational or environmental conditions [8]. State-of-the-art detection frameworks such as CV4CDD-4D [16] can now reliably identify *when* a drift occurs and classify its type with high accuracy. However, they remain descriptive: they leave analysts with unexplained change points that require extensive manual investigation to be actionable [23]. This constitutes a critical *detection-interpretation gap*.

By their nature, event logs capture sequences of activities but rarely the surrounding circumstances, strategic intent, or external pressures that drive process evolution [2, 23]. A sudden increase in approval times, for instance, might reflect

internal process inefficiency, but it might also be the consequence of a new regulatory mandate or a temporary system outage. Such factors are invisible in the event log alone [4, 14]. So far, existing work only addresses this problem partially. The Granger causality framework by Adams et al. [4] relies on statistical correlations between features already present in structured event data but cannot explain drifts triggered by events that leave no structured digital trace; the context-sensitive AI assistant by Brützke et al. [9] enhances accessibility through conversational guidance but does not generate specific, evidence-based explanations for detected drifts; the Process Mining Context Taxonomy [13] provides a crucial structured vocabulary for sense-making, yet has not been operationalized into a tool that actively supports drift interpretation. An approach for systematically linking detected drift signals to potential root causes in unstructured enterprise contexts therefore remains an unaddressed research need.

In this paper, we investigate this gap by pursuing the following research goal: *Design and evaluation of the Concept Drift Explainer (CDE), a prototypical artifact that employs an Large Language Model (LLM)-based agentic approach to explain the underlying causes of concept drifts by systematically linking them to evidence from unstructured enterprise sources.*

The CDE connects the outputs of CV4CDD-4D [16] to unstructured organizational knowledge sources via a multi-agent pipeline that decomposes the explanation task into focused sub-tasks: drift parsing, context retrieval, evidence re-ranking, context mapping, and explanation synthesis. Its central design philosophy is that the role of AI in this analytical task is not to automate final answers, but to accelerate human-led verification by grounding every generated hypothesis in traceable, source-level evidence. This design pattern aims to directly confront the hallucination risk inherent to generative models [11].

A controlled evaluation using real-world event logs from the BPI Challenge 2020 [10] and a synthetic context corpus demonstrated high technical reliability, achieving Recall@2 = 100% and MRR = 0.875. Subsequent expert evaluation confirmed the artifact's practical utility, highlighting, for instance, its positive effects on hypothesis generation and assessment. The paper proceeds as follows: Section 2 reviews the relevant literature; Section 3 outlines the Design Science Research (DSR) methodology; Section 4 describes the CDE artifact; Section 5 reports the evaluation; Section 6 discusses our findings and contributions; and Section 7 states our limitations before concluding the paper.

2 Research Background

Business processes are inherently dynamic and evolve due to internal decisions or external influences. Strategic shifts, regulatory updates, staffing changes, and system replacements continuously alter how work is executed, often without any formal change management record [8]. In process mining, this phenomenon is formalized as *concept drift*: a systematic change in process behavior that occurs while the process is being analyzed [8, 3, 1]. Building on foundational work by Bose et al. [8], four drift types have been established: sudden, gradual, incremental, and recurring. Each drift type describes a distinct pattern of behavioral change over time.

Detecting these changes has matured considerably. Traditional approaches relying on statistical hypothesis testing or trace clustering were limited by fixed assumptions about how drift manifests and were frequently inaccurate on complex, noisy data [20, 16]. The state-of-the-art CV4CDD-4D framework by Kraus and van der Aa [16] addresses this by transforming event logs into image-like representations and applying a computer vision model to identify change points. It is the first supervised approach capable of reliably detecting all four drift types with high accuracy across heterogeneous datasets, representing the current benchmark for concept drift detection. Yet, detection does not yield explanation.

CV4CDD-4D and comparable approaches identify *when* and *what kind* of change occurred, but provide no insight into *why* processes change [16, 4]. This is structurally unavoidable: event logs record sequences of activities but not the organizational conditions surrounding them [23]. A sudden increase in approval cycle times may reflect an internal bottleneck, a new compliance mandate, or a temporary system outage; all equally plausible, but none distinguishable from the log alone. Attempts to close this gap from within the log have remained limited [4, 22]. Adams et al. [4] introduce a Granger causality approach that identifies statistical correlations between process features, but it is necessarily bounded by what is already captured in structured event data and cannot account for triggers that leave no direct digital trace. We refer to this unresolved space as the *detection-interpretation gap*.

Closing this gap requires integrating external, unstructured organizational knowledge into the analysis. The Process Mining Context Taxonomy proposed by Franzoi et al. [13] provides the necessary conceptual vocabulary, distinguishing three levels of contextual influence: process-immediate (e.g., resource or IT changes), organization-internal (e.g., structural changes), and organization-external (e.g., regulatory or environmental changes). This taxonomy identifies what to look for, but does not operationalize the search. LLMs present a promising mechanism for doing so. Their capacity to process heterogeneous unstructured text, reason across semantically diverse sources, and generate human-readable interpretations makes them well-suited for bridging structured drift signals with dispersed organizational context [11, 7, 12]. To date, no existing work has demonstrated an end-to-end pipeline that connects concept drift detector outputs to unstructured enterprise knowledge through LLM-based reasoning [5, 6]. This is the gap the present work addresses.

3 Method

This study adopts DSR as its methodological framework, following the six-phase process model of Peffers et al. [18]. DSR was selected because it directly supports the dual objective of this work: generating prescriptive design knowledge through the construction of a purposeful artifact, while also grounding that artifact in expert-validated problem relevance and practical utility [15]. The scope of this study encompasses a single, complete design-evaluate cycle, positioning the CDE as a proof-of-concept prototype. Evaluation follows the framework of Sonnenberg and vom Brocke [21], combining ex-ante assessments to validate the problem framing and design objectives

prior to implementation with ex-post assessments that demonstrate the artifact's technical performance and practical utility after instantiation.

The CDE integrates three categories of input data. First, five real-world event logs from the BPI Challenge 2020 [10] document a travel expense claim process and collectively contain eight detected concept drifts, identified using the state-of-the-art CV4CDD-4D detector [16]. Second, since no public ground-truth dataset links event log drifts to organizational documents, we constructed a synthetic context corpus of 35 documents, consisting of eight gold documents engineered to explain exactly one drift each, and 27 thematically related noise documents designed to stress-test the system's filtering capabilities (construction procedure detailed in online appendix 3.2). Third, detector outputs in `.csv` and `.json` format serve as the structured bridge between the event logs and the retrieval pipeline. These three streams form the empirical basis for four evaluation activities (Eval 1–4), described in detail in Section 5. We present more details on our methodological approach in Sections 2, 3, and 4 of the online appendix available here.

4 Artifact Description

To bridge the detection-interpretation gap, we designed the *Concept Drift Explainer (CDE)*. The CDE operationalizes this gap as a retrieval and reasoning problem, solved through a chain of specialized agents that progressively transform raw drift detections into traceable, context-grounded explanations. Guided by ten Design Objectives (DOs) derived from the literature and validated ex-ante with six domain experts, the architecture is organized around three clusters: a data and context layer (DO1–2), a reasoning quality layer (DO3–7), and a trust and transparency layer (DO8–10). Full DO specifications are provided in the online appendix. Figure 1 illustrates the architecture of our artifact, which is described in more detail in the following sections.

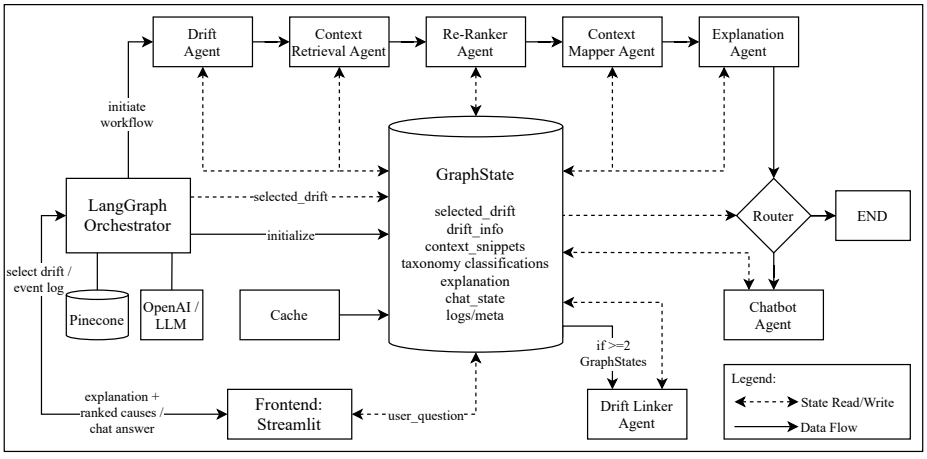

Fig. 1. Agent-based architecture of the CDE.

4.1 Agentic Architecture and State Management

The CDE is operationalized as a directed graph of specialized LLM-based agents, orchestrated via the *LangGraph* framework. This modular design enables explicit state management and controllable hand-offs between reasoning steps, which is critical for ensuring the reliability of process mining insights [7].

Before detailing each component, it is useful to understand the pipeline as a directed flow of progressively refined information. The *Drift Agent* initiates the process by preprocessing raw detector outputs into a semantic anchor: a concise, LLM-generated phrase that captures the essence of the observed process change. The *Context Retrieval Agent* uses this anchor, together with temporally constrained metadata filters, to query a vector index of enterprise documents and assemble a broad candidate set. The *Re-Ranker Agent* sharpens this set by combining vector similarity with semantic specificity scoring, reducing noise and surfacing the most causally plausible evidence. The *Context Mapper Agent* classifies each retained snippet according to the Process Mining Context Taxonomy [13], situating the evidence within a structured socio-technical frame. Finally, the *Explanation Agent* expands each snippet to its full source document, applies a drift-type-specific reasoning chain, and synthesizes a ranked, confidence-scored explanation for the analyst.

The connective tissue of this pipeline is the *GraphState*: a single typed state object that all agents read from and write to in sequence. Each agent reads a well-typed subset of this state, appends its output, and passes the updated object to the next stage. This ensures that every generated explanation is accompanied by its source document and a calibrated confidence score, making all claims directly falsifiable by the analyst.

Beyond the core pipeline, two additional agents support interactive sense-making. The *Chatbot Agent* grounds its responses in the full accumulated state, allowing users to challenge or refine generated hypotheses through natural language follow-up questions. For complex scenarios involving multiple drifts, the *Drift Linker Agent* performs a cross-drift analysis to identify shared root causes or causal dependencies across the log, broadening the analysis from isolated explanations to holistic sense-making. Figure 1 summarizes the architecture of the CDE.

4.2 Implementation

The prototypical instantiation of the CDE uses a hybrid LLM strategy: *GPT-4o* for high-stakes reasoning (ranking) and *GPT-4o-mini* for standard processing tasks. Enterprise context is managed in a *Pinecone* vector database, supporting multimodal ingestion of PDF and PPTX files. The full source code, event logs, and setup documentation are publicly available to support reproducibility [1].

5 Evaluation

Evaluation in DSR serves a dual function: it must validate that the designed artifact fulfills the identified objectives *before* implementation (ex-ante) and demonstrate

[1] `https://anonymous.4open.science/r/concept_drift_explainer-98D4`

its performance and utility *after* implementation (ex-post) [21]. In accordance with this framework, our evaluation was structured as a four-stage process, moving from conceptual validation (Eval 1 and 2) through technical verification (Eval 3) to practical utility assessment (Eval 4).

5.1 Ex-Ante Validation (Eval 1-2)

Prior to implementation, six domain experts (three practitioners and three researchers) validated the relevance of the research problem and the soundness of the ten Design Objectives (DOs) across two ex-ante evaluation rounds (Eval 1–2), combining structured Likert-scale surveys with semi-structured interviews. Experts rated problem importance at 5.8/7 and DO clarity at 6.2/7, confirming both the practical significance of the detection-interpretation gap and the feasibility of the proposed design. Qualitatively, participants emphasized that the problem is most acute in organizations with complex, cross-departmental processes where knowledge is dispersed across siloed systems. Full results of Eval 1-2 are provided in the online appendix.

5.2 Technical Prototype Evaluation (Eval 3)

The technical evaluation assessed the CDE's core capability: given a detected concept drift, does the system retrieve and rank the correct explanatory document from a noisy enterprise corpus? The evaluation was conducted using five real-world event logs from the BPI Challenge 2020 [10], documenting a travel expense claim process at Eindhoven University of Technology that collectively contains eight distinct concept drifts which were detected by CV4CDD-4D [16]. The system's knowledge base comprised a purpose-built synthetic corpus of 35 documents: eight gold documents, each engineered to explain one drift, and 27 thematically related distractors simulating retrieval noise. For each drift, the CDE's task was to place the single correct explanatory document as high as possible in its ranked output. Performance was measured using standard information retrieval metrics: Recall@k and Mean Reciprocal Rank (MRR).

Table 1. Technical Performance Metrics of the CDE Prototype ($n=8$ drifts).

Metric	Value	Interpretation
Recall@1	75% (6/8)	Gold document ranked 1st in 6 of 8 cases.
Recall@2	100% (8/8)	Gold document in top 2 across *all* cases.
MRR	0.875 (of 1.0)	Near-ideal average ranking quality.

We emphasize that this evaluation constitutes a controlled technical benchmark rather than a test of real-world explanatory performance. Because the gold documents were purpose-built to explain their corresponding drift (Sec. 3; online appendix 3.2), the retrieval task is necessarily more tractable than retrieval over authentic, unlabeled enterprise content. The metrics in Table 1 should be read as an upper-bound estimate

of technical feasibility, not as evidence of production-grade retrieval accuracy. The Recall@2 score of 100% is a particularly meaningful result: it demonstrates that the CDE consistently reduces the analyst's search space to at most two candidate documents, effectively transforming what would otherwise be an exhaustive manual review of the entire knowledge base into a targeted validation of one or two specific sources. The MRR of 0.875 reflects the consistently high ranking quality across all evaluated drifts.

5.3 Expert Evaluation (Eval 4)

Eval 4 assessed the CDE's practical utility through a hands-on user study with five participants (P1–P3, R1–R2). Each participant accessed a deployed instance of the CDE and applied it to formulate an explanatory hypothesis for a known incremental drift in the BPI Challenge 2020 international declarations subprocess. Following the task, they completed structured Likert surveys and participated in semi-structured interviews. This scenario-based, task-driven design enables assessment of the artifact's applicability, effectiveness, efficiency, fidelity, and overall impact as perceived by domain experts [21]. As participants worked with a known drift, a public event log, and the synthetic context corpus introduced in Section 3, the findings below should be read as evidence of perceived utility and usability rather than as a measurement of real-world productivity gains.

Overall, participants recognized the CDE as directly addressing a critical analytical pain point: as P2 articulated, *"I'll show you that there's a deviation, but you have to think about the why yourself. That's why this approach is really great."* R1 described the approach as a potential *"game changer"* for the discipline. Overall satisfaction averaged 6.2/7 (SD = 1.17), intention to use the tool in professional practice was equally strong at 6.2/7 (SD = 1.17), and perceived impact on process analysis reached 5.8/7 (SD = 1.94).

The CDE's most immediately recognized benefit was the reduction in manual effort required for root cause investigation. Current practice was described by P2 as an *"extremely difficult and time-consuming"* process of manually searching through dispersed documents without systematic support. The CDE transforms this by automating document retrieval and ranking. As R2 summarized: *"instead of looking at 30 documents, I can concentrate on the one or two most important ones and save a lot of time"*. Participants rated the tool's ability to focus analytical attention at 6.2/7 (SD = 0.98) and its reduction of problem complexity at 6.0/7 (SD = 1.10). When asked to identify the most significant time-saving UI elements, participants selected direct links to source documents (100%), the ranked list of causes (80%), and the causal event timeline visualization (60%). Critically, this efficiency gain is contingent on retrieval quality: if surfaced documents are irrelevant, the system *"would contradict the purpose of the tool, which is to make me faster"* (P3), underlining the direct link between the technical results of Eval 3 and the practical utility assessed here.

Beyond raw efficiency, participants identified a second, arguably more transformative value: the CDE's capacity to democratize root cause analysis by compensating for limited domain knowledge. P2 noted that the tool can *"compensate for skill deficits in employees who may lack the intuition or deeper knowledge for such analyses,"* and P1 successfully formulated a plausible causal hypothesis without deep expertise

in travel expense accounting. The CDE's ability to generate a testable hypothesis received a score of 7.0/7 (SD = 0.00) and the plausibility of generated explanations was rated 6.6/7 (SD = 0.49).

A central finding concerns the mechanism by which analysts construct trust in the CDE's outputs. Baseline trust in explanatory AI stood at 4.0/7 (SD = 1.26) prior to the user study; task-specific trust afterward rose substantially to 5.6/7 (SD = 1.36). This increase was not attributable to confidence in the system's internal reasoning, but mostly to one design feature: source-level traceability. As P1 stated: *"my trust in the results is a 7, because you always have the sources and can form your own picture"*. Participants universally adopted a 'trust but verify' operating model, using the CDE's ranked output as a hypothesis, then validating against the primary source. Crucially, no participant encountered a hallucination during the user study, yet all five nonetheless treated source verification as a non-negotiable step. The tool's value, in this framing, is not to provide infallible answers but to radically accelerate the verification process itself.

This trust dynamic, however, contains a structural tension. R1 observed that a decision-maker is likely to grant more credibility to conclusions reached through lengthy manual research than to identical conclusions derived via an automated tool, even when the underlying evidence is the same. We term this pattern the *Trust Paradox*: while it rests on a single explicit articulation within a small sample (n=5), it aligns with the broader 'trust but verify' stance adopted by all participants, and suggests a residual resistance to AI-mediated reasoning that may be independent of output quality. If confirmed in larger samples, this would imply that the CDE must be designed not only to be correct, but to be legibly correct: its reasoning must be transparent and auditable at every step. Source traceability is the primary mechanism for achieving this, and it reinforces the non-negotiable role of the human analyst as the final decision-maker.

Beyond the overall positive reception, participants surfaced two concrete design limitations that constrain the prototype's current utility. First, the raw numerical confidence score was widely criticized as opaque. R1 described it as a *"metric that I don't understand"* and noted it was *"extremely attackable"* without a scale for interpretation. Participants uniformly relied on source document links rather than the confidence percentage as their primary trust signal, which reinforces our earlier finding on source traceability. This points to a concrete design improvement: replacing raw scores with banded confidence levels (e.g., low/medium/high) accompanied by a brief human-readable rationale. Second, participants identified a missing *drift delta*: the CDE currently explains *why* a change occurred but does not explicitly surface *what* changed in the process flow (e.g., which activities were added or removed). The absence of this target-actual comparison was noted as the most significant functional gap. Notably, despite this omission, participants still rated their ability to identify a likely drift trigger at 6.2/7 (SD = 0.84), suggesting sufficient explanatory fidelity exists within the pipeline and that surfacing the drift delta in the UI represents a high-priority next step rather than a fundamental architectural limitation.

Finally, the expert evaluation surfaced a broader pattern regarding organizational readiness, echoed independently across both the ex-ante (Eval 1–2) and ex-post (Eval 4) rounds. Participants consistently distinguished between the CDE's demonstrated analytical utility and the preconditions required for its deployment, noting that, where

these are met, the artifact offers substantial benefits, while their absence renders even a technically reliable system ineffective. This points to a Maturity Paradox: the organizations that need explanatory tooling most appear least positioned to adopt it.

6 Discussion and Implications

The evaluation establishes the CDE as a viable proof of concept for bridging the detection-interpretation gap identified in Section 2. Where existing approaches reliably identify *when* and *what kind* of process change occurred, the CDE surfaces verifiable, source-linked evidence that supports analysts in forming a plausible hypothesis about *why* it occurred. This is not an incremental extension of existing detection methods; it is a qualitatively different analytical act. The convergence of technical reliability (Recall@2 = 100%, MRR = 0.875) and expert validation confirms both that the pipeline functions as designed and that its outputs are analytically meaningful to the practitioners who would use them.

The CDE advances the literature along three distinct dimensions. First, it operationalizes the Process Mining Context Taxonomy [13] from a conceptual vocabulary into an active analytical component. Through the Context Mapper Agent, every retrieved evidence snippet is classified against the taxonomy's three-level hierarchy (process-immediate, organization-internal, and organization-external), transforming a conceptual framework for sense-making into a functional instrument embedded directly in the analytical workflow. Second, the CDE extends the scope of concept drift explanation beyond log-internal correlates. Prior work, most notably the Granger causality approach of Adams et al. [4], identifies causal relationships within structured event data. The CDE addresses a fundamentally different explanatory challenge: drifts whose triggers leave no direct trace in the event log and are therefore structurally inaccessible to log-internal methods, such as regulatory mandates or organizational restructurings. Third, the CDE instantiates a design pattern for trustworthy explanatory AI in process mining: RAG-grounded, source-traceable outputs that anchor every AI-generated hypothesis to a specific source document [17, 19]. While this pattern is not conceptually specific to concept drift, its transferability to other process mining tasks has not been empirically tested here and remains a direction for future work.

The expert evaluation also surfaces two dynamics that generalize beyond the CDE. The *Trust Paradox* captures a fundamental tension in human-AI collaboration: trust in the system's outputs rose substantially during evaluation (from 4.0/7 to 5.6/7), yet it was constructed not through confidence in the AI's internal reasoning but through the repeated act of source verification. One participant noted that an identical conclusion reached through slow manual analysis would likely be perceived as more credible by decision-makers than one produced through automated inference, suggesting that analytical efficiency and organizational legitimacy may not co-vary. While based on a single explicit observation, this pattern is consistent with the broader verification behavior reported across all participants. This suggests that the human-in-the-loop is not only a safeguard against error, but also a socially necessary component of the legitimation process. If this is confirmed by future research, it will have implications for the design and adoption of any AI-driven explanatory tool [11]. The *Maturity Paradox*

captures an equally structural tension at the organizational level: the organizations that stand to benefit most from explanatory tools like CDE are those least equipped to provide the centralized, governed context corpus on which the CDE depends, while organizations with the data maturity to deploy it may already operate processes formalized enough that contextual drivers of change are more readily known. Taken together, the CDE represents a step toward process mining systems that do not merely surface what happened, but support the effort to understand why it happened.

7 Limitations, Conclusion, and Outlook

As with any research, this work has limitations. First, the evaluation relied on a synthetic context corpus of 35 documents. It served as a controlled proxy, one that is substantially less noisy, redundant, and heterogeneously versioned than real organizational knowledge bases. Therefore, the observed retrieval performance may be optimistic relative to production deployments. Second, expert samples were rather small ($n = 5$–6) and drawn from a single process domain, limiting the generalizability of the qualitative findings across process types and organizational contexts. Third, at the prototype level, the opacity of the raw confidence score represents an unresolved design gap: a metric that participants could not interpret undermines the transparency the system otherwise delivers through source traceability. A concrete next step is to replace it with banded confidence levels (e.g., a low/medium/high classification accompanied by a brief human-readable rationale) transforming an opaque metric into an interpretable signal that supports rather than undermines the verification workflow. Future work should extend the evaluation along three axes: larger and more diverse expert samples, event logs and organizational settings beyond the single travel-expense domain studied here, and alternative LLM architectures and providers to test the robustness of the retrieval and reasoning pipeline beyond the GPT-4o family used in this prototype.

Notwithstanding these limitations, the CDE provides initial evidence that the detection-interpretation gap in concept drift analysis can be narrowed through retrieval-augmented, source-traceable hypothesis generation. Confirming this at production scale remains an open empirical question. The design pattern introduced here offers a principled response to the hallucination and opacity challenges that constrain the deployment of generative AI in analytical workflows. Whether it transfers to process mining tasks beyond concept drift explanation remains to be tested in future work. The challenge now shifts from demonstrating feasibility to navigating the preconditions for organizational adoption: data provisioning, governance, and the maturity required to sustain a living context corpus. The long-term vision is a semi-autonomous process management loop in which the system not only explains detected drifts but proposes corrective actions for human approval, helping organizations not only understand process dynamics, but act on them.

References

[1] van der Aalst, W.M.P.: Process Mining: Data Science in Action. Springer Berlin, Heidelberg, 2nd edn. (2016)

[2] van der Aalst, W.M.P.: Process Mining: A 360 Degree Overview. In: van der Aalst, W.M.P., Carmona, J. (eds.) Process Mining Handbook, pp. 3–34. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-08848-3_1

[3] van der Aalst, W.M.P., Adriansyah, A., et al.: Process Mining Manifesto. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) Business Process Management Workshops. pp. 169–194. Springer, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28108-2_19

[4] Adams, J.N., van Zelst, S.J., Rose, T., van der Aalst, W.M.P.: Explainable concept drift in process mining. Information Systems **114** (Mar 2023). https://doi.org/10.1016/j.is.2023.102177

[5] Barbieri, L., Stroeh, K., Madeira, E.R.M., van der Aalst, W.M.P.: An LLM-Based Q&A Natural Language Interface to Process Mining. In: Delgado, A., Slaats, T. (eds.) Process Mining Workshops. pp. 5–17. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-82225-4_1

[6] Berti, A., Kourani, H., Hafke, H., Li, C.Y., Schuster, D.: Evaluating Large Language Models in Process Mining: Capabilities, Benchmarks, and Evaluation Strategies. In: van der Aa, H., Bork, D., Schmidt, R., Sturm, A. (eds.) Enterprise, Business-Process and Information Systems Modeling, LNBIP, vol. 511, pp. 13–21. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-61007-3_2

[7] Berti, A., Maatallah, M., Jessen, U., Sroka, M., Ghannouchi, S.A.: Re-Thinking Process Mining in the AI-Based Agents Era (Aug 2024). https://doi.org/10.48550/arXiv.2408.07720

[8] Bose, R.P.J.C., van der Aalst, W.M.P., Žliobaitė, I., Pechenizkiy, M.: Handling Concept Drift in Process Mining. In: Mouratidis, H., Rolland, C. (eds.) Advanced Information Systems Engineering. pp. 391–405. Springer, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21640-4_30

[9] Brützke, P., Killewald, R., Franzoi, S., vom Brocke, J.: AI-Assisted Process Mining for Context-Sensitive Analysis Support. ECIS 2025 Proceedings **33** (2025)

[10] van Dongen, B.: BPI Challenge 2020 (Mar 2020). https://doi.org/10.4121/UUID:52FB97D4-4588-43C9-9D04-3604D4613B51, 4TU.Centre for Research Data

[11] Feuerriegel, S., Hartmann, J., Janiesch, C., Zschech, P.: Generative AI. Business & Information Systems Engineering **66**(1), 111–126 (Feb 2024). https://doi.org/10.1007/s12599-023-00834-7

[12] Franzoi, S., Delwaulle, M., Dyong, J., Schaffner, J., Burger, M., vom Brocke, J.: Using Large Language Models to Generate Process Knowledge from Enterprise Content. In: Gdowska, K., Gómez-López, M.T., Rehse, J.R. (eds.) Business

Process Management Workshops. pp. 247–258. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-78666-2_19

[13] Franzoi, S., Hartl, S., Grisold, T., van der Aa, H., Mendling, J., vom Brocke, J.: Explaining process dynamics: a Process Mining Context Taxonomy for sense-making. Process Science **2**(1) (Mar 2025). https://doi.org/10.1007/s44311-025-00008-6

[14] Grisold, T., van der Aa, H., Franzoi, S., Hartl, S., Mendling, J., vom Brocke, J.: A Context Framework for Sense-making of Process Mining Results. 6th International Conference on Process Mining, ICPM 2024, Kgs. Lyngby, Denmark, October 14-18, 2024 pp. 57–64 (2024). https://doi.org/10.1109/ICPM63005.2024.10680665

[15] Hevner, A.R., March, S.T., Park, J., Ram, S.: Design Science in Information Systems Research. MIS Quarterly **28**(1), 75–105 (2004). https://doi.org/10.2307/25148625

[16] Kraus, A., van der Aa, H.: Machine learning-based detection of concept drift in business processes. Process Science **2**(1) (Apr 2025). https://doi.org/10.1007/s44311-025-00012-w

[17] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (Apr 2020). https://doi.org/10.48550/arXiv.2005.11401

[18] Peffers, K., Tuunanen, T., Rothenberger, M.A., Chatterjee, S.: A Design Science Research Methodology for Information Systems Research. Journal of Management Information Systems **24**(3), 45–77 (Dec 2007). https://doi.org/10.2753/MIS0742-1222240302

[19] Rai, A.: Explainable AI: from black box to glass box. Journal of the Academy of Marketing Science **48**(1), 137–141 (Jan 2020). https://doi.org/10.1007/s11747-019-00710-5

[20] Sato, D.M.V., Freitas, S.C.d., Barddal, J.P., Scalabrin, E.E.: A Survey on Concept Drift in Process Mining. ACM Computing Surveys **54**(9), 1–38 (2021). https://doi.org/10.1145/3472752

[21] Sonnenberg, C., vom Brocke, J.: Evaluations in the Science of the Artificial – Reconsidering the Build-Evaluate Pattern in Design Science Research. In: Peffers, K., Rothenberger, M., Kuechler, B. (eds.) Design Science Research in Information Systems. Advances in Theory and Practice. pp. 381–397. Springer, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29863-9_28

[22] Yeshchenko, A., Di Ciccio, C., Mendling, J., Polyvyanyy, A.: Comprehensive Process Drift Detection with Visual Analytics. In: Laender, A.H.F., Pernici, B., Lim, E.P., de Oliveira, J.P.M. (eds.) Conceptual Modeling. pp. 119–135. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-33223-5_11

[23] Zimmermann, L., Zerbato, F., Weber, B.: What makes life for process mining analysts difficult? A reflection of challenges. Software and Systems Modeling **23**, 1345–1373 (2024). https://doi.org/10.1007/s10270-023-01134-0