Implicit Feedback in EdTech: Technical State of the Art

The interpretation of implicit feedback signals—clickstream data, dwell time, and behavioral patterns—has emerged as the foundation of modern adaptive learning systems. **Deep learning architectures now achieve 80-85% AUC in predicting learning outcomes from implicit signals alone**, with transformer-based knowledge tracing models (SAKT, AKT) and graph neural networks (GKT) outperforming traditional approaches by significant margins. However, substantial challenges persist: time-on-task estimation methods alone can cause **23% variance in explained outcome variance** depending on methodology choice, (learning-analytics) highlighting how preprocessing decisions fundamentally shape analytical validity. This report synthesizes peer-reviewed research from 2020-2025 on clickstream analysis, dwell time normalization, multi-signal fusion, and critical implementation considerations including privacy and interpretability.

Clickstream architectures have converged on two dominant standards

Modern EdTech platforms capture implicit feedback through standardized event taxonomies, with **xAPI (Experience API)** and **IMS Caliper Analytics** representing the primary data collection frameworks. xAPI employs an Actor-Verb-Object triple structure in JSON format, optimized for single-application freeform logging across video interactions, reading behaviors, simulations, and offline activities. IMS Caliper, by contrast, uses JSON-LD with stricter semantic enforcement through defined Metric Profiles (Annotation, Session, Assessment, Grading), making it better suited for cross-application aggregation where institutions need standardized event semantics.

The distinction matters: xAPI excels at capturing nuanced micro-behaviors within a platform, while Caliper provides better interoperability for institutional-level analytics. Recent convergence efforts between ADL and 1EdTech aim to bridge these frameworks, with xAPI 2.0 under ISO/IEC standardization. Event taxonomies typically capture navigation events (page views, module access), content interactions (video play/pause/seek with timing), assessment events (quiz attempts, problem-level responses), and social behaviors (forum posts, peer interactions). The Open University Learning Analytics Dataset (OULAD), a widely-used benchmark with **32,593 students and 10.6 million VLE interactions**, organizes events across 12 learning site categories that serve as a de facto taxonomy standard.

Sequential pattern mining reveals learning strategies

The extraction of meaningful patterns from clickstream sequences employs several algorithmic families. **PrefixSpan** (MOOC List) (Prefix-projected Sequential pattern mining) dominates educational data mining applications due to its memory efficiency with large datasets, though it struggles with smaller educational samples. GSP (MOOC List) (Generalized Sequential Patterns) and cSPADE (MOOC List) offer alternatives optimized for horizontal and vertical data formats respectively. More recent work addresses MOOC-specific challenges: Song et al. (2022) developed algorithms mining sequential patterns with flexible constraints for enrollment sequences, while FAST-USP specifically targets "unexpected" patterns that deviate from typical learning behaviors. (IEEE Xplore)

Data preprocessing proves critical to pattern quality. Zhang and Paquette (2023) identify five essential preprocessing steps: **filtering** irrelevant automatic events, **collapsing** consecutive identical actions, **contextualizing** with temporal markers, **abstraction** into higher-level activity categories, and **breaking** sequences into meaningful session units. Process mining tools—particularly **ProM 6.x** using XES format logs—enable discovery of actual learning paths versus designed curricula, with algorithms like Inductive Visual Miner and Fuzzy Miner revealing bottlenecks and conformance deviations. The Trace-SRL framework specifically maps clickstream traces to self-regulated learning processes through strategic clustering and micro-level processing.

Feature engineering for clickstream data spans handcrafted temporal features (weekly click aggregations, deadline-relative timing, inter-session gaps) and learned representations. Recent architectures use LSTM auto-encoders for fixed-length embeddings (ResearchGate) and graph convolutional networks for video viewing behaviors. The ClickTree method, placing in the 2024 EDM Cup, combines three-level feature extraction (problem, assignment, student levels) with CatBoost classification to achieve **~79% AUC**. (Educationaldatamining) Processing architectures increasingly adopt Lambda patterns combining batch layers (Hadoop/Spark) with speed layers (Kafka Streams, Flink) for real-time intervention triggers alongside historical analysis.

Dwell time normalization critically affects analytical validity

A foundational study by Kovanović et al. (2015) systematically evaluated **15 time-on-task estimation strategies**, revealing that methodology choice alone accounts for R^2 variance differences up to 0.23 across performance measures. (learning-analytics) The optimal strategy—replacing actions exceeding 10 minutes with per-user averages (+10ev)—consistently outperformed naive approaches, yet timeout interval choices in literature range from 80 seconds (Baker's off-task behavior threshold) to 60 minutes (Wise et al.'s inactivity indicator). This methodological heterogeneity undermines cross-study comparability and replication efforts.

Content complexity adjustment requires integrating multiple signals. Traditional readability formulas (Flesch-Kincaid, Coleman-Liau) provide baseline adjustments, but NLP-enhanced approaches like **CAREC** and **Coh-Metrix L2** incorporate lexical diversity, syntactic complexity, cohesion features, age-of-acquisition metrics, and character entropy. Normalization formulas typically divide raw dwell time by content-length factors calibrated using reading speed baselines. Brysbaert's meta-analysis of 190 studies establishes **238 WPM for silent non-fiction reading** as a reference, (ResearchGate) though learning tasks reduce this to 100-200 WPM and memorization to under 100 WPM. (WordsRated) Individual calibration approaches establish per-user baselines from initial interactions and progressively update reading profiles.

Discriminating attention from distraction increasingly relies on machine learning. Betto et al. (2023) combine Eye Area Ratio, gaze angle, and posture features from facial landmark detection, achieving **95% F1-score** using Random Forest and XGBoost classifiers. VR-based eye tracking with personalized models reaches 98.88% accuracy for three-level distraction classification. Threshold-based discrimination flags dwell times below 0.15 seconds per word as scanning and times exceeding session-specific thresholds as potential distraction. Dwell time distributions consistently follow **log-normal patterns**, (Wikipedia) requiring log-transformation before parametric analysis; outlier detection using MAD (Median Absolute Deviation) proves more robust than mean/SD methods for these skewed distributions.

Fusion architectures balance modality interaction against flexibility

Multi-signal fusion for learning preference inference operates across three architectural paradigms. **Early fusion** combines raw features before model processing through concatenation or shared embeddings, capturing cross-modal correlations at the lowest level but suffering from curse-of-dimensionality and sensitivity to missing modalities. (GeeksforGeeks) **Late fusion** trains independent models per signal type and combines predictions through ensemble voting, weighted averaging, or stacking—providing modularity and robustness to missing data but losing fine-grained cross-modal interactions. (GeeksforGeeks) **Intermediate fusion** processes features separately, then fuses in latent space before classification, offering a principled middle ground.

Attention mechanisms have transformed educational signal fusion. Self-Attentive Knowledge Tracing (SAKT), the first transformer-based knowledge tracing model, weights past interactions for predicting knowledge component mastery while running an order of magnitude faster than RNN approaches. AKT (Attentive Knowledge Tracing) adds monotonic attention modules modeling knowledge decay, achieving **~0.78 AUC** on ASSISTments datasets. Cross-modal attention architectures use Query-Key-Value mechanisms between modalities, with attention bottleneck designs restricting cross-modal information flow through defined channels to reduce computational complexity while maintaining performance. (arXiv) The KANFormer architecture integrates multi-head self-attention with Kolmogorov-Arnold Networks for processing student demographic, academic, and engagement features simultaneously.

Uncertainty quantification enables calibrated predictions

Uncertainty-aware fusion addresses the fundamental challenge that different signal modalities carry different reliability levels. Monte Carlo Dropout, derived from Gal and Ghahramani's framework, enables epistemic uncertainty quantification in knowledge tracing models, with larger predictive uncertainty empirically aligning with incorrect predictions. Bayesian neural networks maintain prior distributions over weights, updating posteriors through observed data and decomposing uncertainty into reducible (model) and irreducible (data) components.

Evidential deep learning offers particularly promising capabilities for educational contexts. The approach maps features to mass functions using Dempster-Shafer theory, with contextual discounting coefficients quantifying source reliability relative to each class. (ACM Digital Library) (arXiv) Dual-level Deep Evidential Fusion (DDEF) integrates at both Basic Belief Assignment and multimodal levels, demonstrating enhanced accuracy and—critically—providing interpretable contribution metrics per modality. Huang et al. (2024) show that learned reliability coefficients automatically assess signal quality without explicit labels, enabling the system to weight high-quality signals more heavily during fusion.

Temporal alignment of heterogeneous signals employs Dynamic Time Warping (DTW) and its variants. (Wikipedia) Standard DTW has $O(N^2)$ complexity, while FastDTW achieves $O(N)$. For educational data, **Canonical Time Warping** combines CCA-based feature space transformation with DTW alignment, and GromovDTW uses time series self-similarities for handling signals with fundamentally different measurement spaces. (RTavenar) The WHEN architecture (KDD 2023) addresses intra-sequence non-stationarity through

wavelet attention and inter-sequence asynchronism through DTW transformed into attention form, demonstrating state-of-the-art results on sequence alignment tasks.

Graph and sequence models dominate knowledge tracing

Deep learning architectures for educational signal processing have evolved substantially since Deep Knowledge Tracing (DKT) introduced LSTMs to the field in 2015. Graph-based approaches now achieve superior performance: **Graph-based Knowledge Tracing (GKT)** reformulates the problem as time-series node-level classification over knowledge structure graphs, (ACM Other conferences) reaching **0.82 AUC** while providing improved interpretability over black-box alternatives. GIKT uses graph convolutions to model question-knowledge point correlations, while HHSKT employs heterogeneous graph neural networks with multi-level summarization of knowledge structures. (ScienceDirect)

Hybrid architectures combine the strengths of different paradigms. SSKT integrates LSTM for sequential encoding with transformer attention for global dependencies, explicitly grounded in Ausubel's cognitive theory. GELT combines GNN interpretability with transformer modeling through energy-saving attention mechanisms. For sequential signals, attention-based LSTM achieves **89.6% accuracy** in student performance prediction, substantially outperforming Random Forest (81%) and basic ANN (85%) baselines. (RSIS International) The pyKT library provides standardized implementations of 10+ deep learning knowledge tracing models with rigorous evaluation protocols, (arXiv) addressing the reproducibility crisis that had plagued earlier research.

Privacy-preserving methods achieve near-parity with centralized learning

The DEFLA framework, presented at LAK 2025, represents the first differential privacy framework specifically designed for learning analytics, demonstrating on OULAD that reasonable AUC can be maintained at meaningful privacy budget ε values. DP-TabNet achieves **80% accuracy at ε=0.7** compared to 84% non-private—a modest 4% reduction for strong privacy guarantees. Knowledge tracing under differential privacy, with user-level sequence protection, maintains competitive AUC on ASSISTments and EdNet datasets using Rényi DP, Gaussian DP, and numerical composition accounting.

Federated learning architectures enable model training without centralizing sensitive student data. (Springer) The FecMap model implements Local Subspace Learning with Multi-layer Privacy Protection, achieving competitive performance to centralized learning on higher education datasets. (Hep) The FLAME metric provides a novel framework for assessing privacy-performance trade-offs, finding that federated approaches achieve **less than 5% accuracy drop** versus centralized training typically. Enhanced privacy combines federated architectures with secure aggregation protocols and CKKS homomorphic encryption for encrypted computation. (IACR) Privacy-preserving synthetic data generation using CTGAN and TVAE, combined with DP mechanisms, protects against linkage attacks while maintaining utility for downstream analytics, though utility decreases sharply below ε=1.0.

Cold-start and transfer learning enable cross-course generalization

Cold-start scenarios—new students, courses, or items without training data—remain fundamental challenges for

implicit feedback systems. Empirical evaluation shows that DKT, DKVMN, and SAKT all struggle initially under cold-start, with SAKT showing higher initial accuracy but plateauing, while DKVMN excels in early predictions due to its memory mechanisms. (arXiv) Transfer learning approaches offer solutions: course-agnostic models avoid course-specific features, enabling naive transfer to new courses with performance comparable to course-specific Bayesian Knowledge Tracing and Performance Factors Analysis. (DeepAI)

Domain adaptation techniques from deep learning translate effectively. **CORAL loss** aligns second-order statistics between source and target domains, preventing overfitting to source features and improving transfer between related MOOC courses. The ACKT framework employs mixture-of-experts networks with adversarial discriminators for cross-disciplinary transfer, clustering student knowledge states as transfer bridges. OCLCKT uses parallel discipline-independent knowledge state retrievers with contrastive learning across disciplines. Most promisingly, the CLST framework aligns large language models as knowledge tracers, achieving **24.52% improvement** over baselines with only 8 training students—a dramatic reduction in data requirements for new course deployment.

Interpretability methods reveal model decision processes

Post-hoc explanation techniques enable understanding of black-box implicit feedback models. **SHAP** (SHapley Additive exPlanations) provides both local and global explanations through game-theoretic feature attribution, with TreeExplainer running 100x faster than KernelExplainer for tree-based models. In student adaptability prediction, SHAP analysis revealed "Class Duration" as the most significant feature with mean SHAP value 0.175. LIME perturbs input features and fits local linear models, though comparative studies show considerable variability in feature importance interpretation across explainer methods.

Attention visualization in self-attentive knowledge tracing reveals which past interactions influence predictions, exposing temporal patterns in learning. (PLOS) AKT's Rasch model embeddings for item difficulty enable interpretation of attention weights as knowledge concept relationships. Graph-based architectures provide inherently more interpretable structures, with GELT designed specifically to uncover skill-question relationships through graph embeddings. Knowledge-Based Artificial Neural Networks (KBANNs) integrate structured educational knowledge as constraints, producing explanations more aligned with educational principles than purely data-driven alternatives. Counterfactual explanations—showing minimum changes needed to alter predictions—prove particularly valuable for policy formulation and intervention targeting.

Benchmark datasets and evaluation protocols require careful selection

The field has consolidated around several benchmark datasets with known characteristics. ASSISTments datasets range from 4,217 students with 346,860 interactions (2009) to 27,000+ students with 2.54 million interactions (2012), all with knowledge component annotations. OULAD provides 32,593 students across 22 module presentations with demographics, registration, assessments, and 10.6 million VLE interactions. **XES3G5M** (NeurIPS 2023) offers the largest KC set in mathematics with 865 knowledge components, 7,652 questions, and hierarchical KC routes enabling fine-grained analysis. EdNet represents the largest scale at 784,309 students with 131 million interactions.

Evaluation protocols critically affect reported performance. The pyKT benchmark identified label leakage and improper train/test splits as causing performance inflation across the literature. Recommended protocols include **student-wise splitting** (train on some students, test on others) for realistic deployment scenarios, temporal/chronological validation ensuring training data precedes test data, and blocked cross-validation for data with temporal dependencies. Standard metrics include AUC-ROC for classification tasks (0.80-0.90 considered good), RMSE for continuous predictions, and precision@k/NDCG for recommendation-style evaluations. Critical research gaps remain: cross-institutional generalization is rarely validated, algorithmic fairness across demographic subgroups remains understudied, and optimal privacy budget selection remains application-dependent.

Conclusion

The technical infrastructure for implicit feedback interpretation in educational technology has matured substantially, with standardized data collection (xAPI/Caliper), validated preprocessing methods, and deep learning architectures achieving 80%+ AUC in predicting learning outcomes. **Three insights stand out**: First, methodological choices in time-on-task estimation and dwell time normalization fundamentally shape analytical conclusions—the +10ev strategy with content-complexity adjustment should be standard practice. Second, evidential deep learning and uncertainty-aware fusion provide not just improved accuracy but interpretable reliability metrics essential for high-stakes educational decisions. Third, privacy-preserving techniques now achieve near-parity with centralized approaches, enabling deployment in institutional contexts with strict data governance requirements. The field's next frontier lies in cross-institutional transfer learning and fairness validation across demographic subgroups—areas where current evidence remains sparse despite significant practical importance.