Stockholm School of Economics
Department of Economics
5350 Master's Thesis in Economics
Spring 2026

Fingertoppskänsla or Foundation Model? Human-Engineered LSTM Features vs. MOMENT in Early-Stage Music Rights Acquisition

Dávid Miženko (42847) and Petter Hallqvist (42843)

Abstract: Music rights are an emerging alternative asset class with skewed payoffs. We treat early-stage rights acquisition as an ex-ante selection problem under acute data scarcity, evaluating whether a fine-tuned MOMENT foundation model outperforms a feature-engineered LSTM at converting limited streaming history into profitable acquisitions. We contribute to the empirical asset pricing literature on machine learning for selection in illiquid alternative markets, and ask the fundamental question - should we just let the models learn?

Keywords: music rights valuation, viral prediction, probability calibration, time series foundation models, asset pricing, portfolio decisions, machine learning

JEL: C45, C53, G11, G12

Supervisor: Rickard Sandberg
Date submitted: 13th of May 2026
Date examined: 28th of May
Discussants: Anton Ottosson, Emma Bergman Karlsson
Examiner: Álvaro Jañez Garcia

Acknowledgements

We would like to thank all the people at RUN for making this project possible, we learned a lot and it was a pleasure working with all of you! Special thanks to Oscar Nordström and Jonathan Westerdahl for ideas of varying quality, great answers to some of our questions and any resources we needed. We are also grateful to our supervisor Rickard Sandberg for all input and help with modelling and writing.

AI Disclosure

This thesis was prepared with the assistance of two AI tools: Claude AI and OpenAI's ChatGPT. Their use was confined to a narrow set of auxiliary tasks. Claude AI was used to write the plotting code that generates the graphs and figures displaying model outputs. ChatGPT supported thematic literature discovery, with every suggestion verified against the underlying sources, and was used to transcribe mathematical expressions into LaTeX rather than typing them by hand. At no point did either tool function as a source of substantive content.

Contents

1 Introduction

Streaming has converted music rights from an obscure royalty stream into a tradable asset class, attracting institutional and private capital over the past decade. Global recorded music revenue exceeded $28 billion in 2023, with streaming generating more than two-thirds of the total (IFPI, 2024). Bob Dylan's 2020 sale of his publishing catalogue to Universal Music Group was one of several transactions that drew mainstream financial attention to the space (Universal Music Group, 2020). This thesis is relevant to pricing-driven investors who lack the capital to influence consumption directly and whose edge is forecasting future cash flows more accurately than competing buyers.

Valuing individual songs is harder than valuing catalogues. For songs with sufficient history, valuation is reasonably tractable: most tracks settle into a predictable decay pattern after their initial peak, and standard discounted cash flow methods perform adequately once that pattern is observable. The hard problem is new releases. Only a few weeks of streams are available, and the firm that values the song faster typically wins the deal — a speed-versus-accuracy trade-off that defines the operational tension in this market. Early viral identification sits at the extreme of this trade-off, where the largest potential returns and the greatest forecasting uncertainty coincide. The decision structure shares features with the real-options framework of Dixit and Pindyck (1994): irreversible capital commitment under high uncertainty, with a competitive race to act.

Early daily streaming trajectories may contain information about momentum, acceleration, and breakout potential that static song features cannot capture. Long-horizon popularity is partly irreducible because it depends on social-feedback dynamics rather than intrinsic track quality alone (Salganik, Dodds, & Watts, 2006), but the relevant question for an early-stage investor is narrower: whether short-horizon trajectory shapes carry exploitable signal over realistic investment windows. We focus on songs that go viral early in their life cycle, since these patterns are more comparable across songs and therefore amenable to large-scale labelling.

Within music, Hit Song Science has applied machine learning to popularity prediction for over a decade, but most of this work has used static features — audio characteristics, metadata, artist popularity — or chart-membership labels, rather than modelling streaming trajectories directly. Modelling the trajectory introduces a representation question: how should a short, sparse time series be encoded before it reaches the predictive model? Two approaches dominate. The first feeds domain-engineered features (moving averages, growth rates, day-of-week patterns) into a recurrent model. The second applies time-series foundation models pre-trained on large diverse sequence corpora directly to raw trajectories. This feature-engineering versus representation-learning question has been substantially settled in computer vision and NLP in favour of learned representations, but remains open for sparse domain-specific time series like music streaming, where no direct analog exists in the foundation-model pre-training corpus.

This is the central question of the thesis:

> *In early-stage music streaming prediction, do foundation-model representations of raw streaming trajectories outperform a recurrent model built on domain-engineered features, and does any performance advantage translate into better portfolio investment outcomes?*

We instantiate each side with a representative model: a Long Short-Term Memory network (LSTM) for the feature-engineered approach, and MOMENT (Goswami et al., 2024) — an open time-series foundation model designed for transfer to downstream forecasting and classification tasks — for the foundation-model approach. We answer the central question in four sequential stages:

1. **Classification:** Does model choice affect viral candidate classification from the first 35 days of streaming data?
2. **Valuation:** Does the LSTM-selected or MOMENT-selected portfolios produce stronger realised investment outcomes?
3. **Portfolio construction:** Do Kelly-inspired sizing rules turn these signals into stronger investment performance than simpler heuristic allocation?
4. **Economic baseline:** Is it better to forecast revenue through peak and decay projection, or to predict total future streams directly?

The thesis makes two contributions. First, it builds an end-to-end pipeline from early streaming data to investment decisions — classification, quantile-based trajectory forecasting, fair market price modelling, IRR and ROI estimation, and backtested portfolio construction in a single integrated framework. Second, it extends prior work on music popularity prediction by comparing foundation-model representations with feature-engineered approaches on early streaming trajectories.

This work was developed in collaboration with RUN, a music rights investment company facing exactly the valuation problem described above. RUN provided access to daily streaming data from Luminate, market insight into the deal flow, and the investment logic that shaped the backtesting design.

The remainder of the thesis is organised as follows. Section 2 reviews the literature on music valuation, streaming forecasting, and machine learning approaches to sequential prediction. Section 3 describes the data and labelling methodology. Section 4 presents the predictive modelling framework. Section 5 operationalises the valuation outputs in a walk-forward portfolio backtest. Section 6 reports empirical results across the four stages. Section 7 discusses implications, limitations, and next steps. Section 8 concludes.

2 Literature Review

Music royalties are income streams paid to musicians, songwriters, and other rights holders when their music is commercially used, including through streaming, broadcasting, public performance, and synchronisation in films, television, and advertisements. Traditionally, these rights and associated revenues were primarily managed by record labels and music publishing companies. However, changes in music consumption patterns, most notably the rise of digital streaming, have significantly altered how music generates value. Due to increasing data availability in music and the finance industry's search for alternative assets beyond traditional stocks and bonds, music rights are increasingly viewed as investable royalty-generating assets (Davies, Cole, & Turner, 2022).

Music copyrights are typically valued using methods similar to those applied to other financial assets. In particular, the literature highlights two dominant approaches: discounted cash flow (DCF) valuation, which estimates present value based on expected future cash flows, and relative valuation methods based on comparable transactions or market multiples (Damodaran, 2012, p. 727). A complicating factor is that music rights also derive value from intangible factors such as cultural relevance, artist reputation, and symbolic appeal. These factors can influence future monetisation opportunities and may vary across genre, rights type, artist language, geographic exposure, and broader industry conditions (Cuntz, Kos, & Soriano Valdes, 2025; Stoikov & Kosyuk, 2022).

Today, the majority of recorded music revenue comes from streaming, meaning that projecting future streams is central to pricing music assets (IFPI, 2024). As data availability and reporting frequency increase, music valuation is moving beyond static financial history toward richer indicators of future performance (Stoikov, Singla, Cetin, & Cendra Villalobos, 2026). Forecasting demand for cultural goods remains difficult because popularity depends not only on intrinsic quality but also on social influence dynamics. In a landmark experimental study, Salganik et al. (2006) created an artificial music market and found that stronger social influence increased both inequality and unpredictability of outcomes, suggesting that music success is partly path-dependent and socially reinforced.

Prior research in Hit Song Science suggests that machine learning can identify patterns associated with commercial music success. Middlebrook and Sheik (2019) predict Billboard Hot 100 success using Spotify audio features and metadata, finding that Random Forest models outperform logistic regression, a simple neural network, and support vector machines. Similarly, Dimolitsas, Kantarelis, and Fouka (2023) use Spotify-derived audio features and report that Random Forest and SVM classifiers achieve approximately 86% accuracy. However, these studies rely primarily on static song-level attributes rather than post-release consumption dynamics. This motivates the use of time-series classification models, which treat streaming observations as ordered market signals that aggregate listener behaviour, social diffusion, promotion, genre effects, and latent song quality.

More closely related to this thesis, Soares Araujo, Pinheiro de Cristo, and Giusti (2019) use

historical Spotify chart information to predict whether songs will enter the Spotify Top 50 Global ranking up to two months ahead, reporting AUC values above 0.80. This suggests that platform-level popularity dynamics contain useful predictive information about future success. However, their framework remains a classification problem based on chart membership rather than a continuous forecasting and valuation problem. In contrast, this thesis models daily streaming trajectories directly and translates predicted future consumption into royalty cash flows and portfolio decisions.

Long Short-Term Memory (LSTM) networks, introduced by Hochreiter and Schmidhuber (1997), address the vanishing gradient problem of traditional recurrent networks through gated memory cells that selectively retain relevant historical information. This makes them well suited to streaming trajectories, where daily listener counts evolve through momentum, decay, retention, and occasional re-acceleration. Static cross-sectional models can identify songs with characteristics associated with success, but they are less suited to modelling how early consumption patterns unfold over time.

Evidence from adjacent forecasting domains supports the use of deep learning for sequential prediction. Reviewing 187 financial forecasting studies, Giantsidi and Tarantola (2025) find that recurrent architectures, particularly LSTMs, remain widely used in noisy and nonlinear time-series environments. They also highlight the increasing use of hybrid architectures and multi-source feature integration to improve predictive robustness. Although their focus is financial markets, the methodological parallel is relevant because future royalty cash flows similarly depend on complex and evolving temporal dynamics.

Recent forecasting literature has also moved toward multi-horizon and probabilistic deep learning architectures. Lim, Arik, Loeff, and Pfister (2020) introduce the Temporal Fusion Transformer, an attention-based model for interpretable multi-horizon forecasting with static covariates, observed historical inputs, known future inputs, and quantile outputs for prediction intervals. This is relevant because royalty valuation requires not only a single forecast of future streaming demand, but an estimated future trajectory under uncertainty. Although this thesis uses a simpler recurrent forecasting framework, the Temporal Fusion Transformer illustrates the broader direction of the literature toward models that combine multi-step forecasting, uncertainty estimation, and interpretability.

The link between prediction and investment decisions is also visible in financial time-series applications. Fjellström (2022) frames stock movement prediction on OMX30 constituents as a binary classification problem and shows that LSTM-derived signals can be translated into portfolios with higher cumulative returns and lower volatility than benchmark portfolios. Zeng et al. (2023) combine convolutional layers with Transformers to capture both local short-term patterns and longer-range dependencies in intraday stock prices. Neither study concerns music directly, but both support the broader argument that sequence models are most economically relevant when forecasts are converted into investment decisions.

Music-rights returns are highly right-skewed: a small number of successful assets can account for a disproportionate share of portfolio returns, while many acquisitions may underperform.

This resembles the payoff structure discussed in venture-capital and power-law return settings (Cochrane, 2005; Gabaix, 2009). Such skewness motivates probabilistic valuation rather than reliance on a single point forecast. In this thesis, uncertainty is incorporated through quantile-based revenue projections, while acquisition cost is modelled through a parsimonious flat-then-decay fair market price construction in the DCF tradition (Damodaran, 2012). The assumed post-peak decay structure can also be interpreted as a simplified post-peak analogue to diffusion-style demand models such as Bass (1969).

Taken together, this literature motivates the modelling strategy of this thesis. Existing Hit Song Science studies show that machine learning can classify commercial success, but they often rely on static song-level features or chart membership labels. The forecasting literature, by contrast, provides tools for modelling sequential dynamics and uncertainty. This thesis combines these ideas by using early daily streaming trajectories to forecast future consumption and translate those forecasts into royalty valuation and portfolio selection decisions. The portfolio perspective is particularly relevant because returns from newly released songs may be highly skewed, with a limited number of breakout successes potentially offsetting weaker acquisitions.

3 Data, Labelling and Modelling Assumptions

The data used in this thesis are sourced from Luminate (formerly Nielsen Music / MRC Data), an industry-standard measurement platform for music consumption. Luminate aggregates verified streaming, sales, airplay, and download data from more than 500 authorised data partners, including major digital service providers (DSPs), and its data are used for industry charts and commercial decision-making across the music sector.

This thesis uses daily streaming consumption data accessed through a Snowflake data warehouse environment. Each track is identified by a unique internal identifier (`MR_ID`), and daily streaming volumes are reported as total streams across reporting DSPs for a given market. The data are subject to a known three-day reporting lag, which is incorporated into the observation-window assumptions and the decision-timing assumptions in Section 4.5.

3.1 Dataset Construction

The modelling dataset is constructed in three steps. First, a random sample of 50 million tracks is drawn from the full Luminate database and stored as `SAMPLE_50M_TS`. Second, tracks whose all-time peak streaming day occurs between day 45 and day 180 post-release are extracted into `SAMPLE_50M_TS_PEAK_45_180`. Third, a consolidated SQL query, shown in Appendix [A], applies the viral-label criteria below and assembles the final three-class modelling structure: viral positives, hard negatives drawn from the peak-filtered subset, and easy negatives drawn from the broader 50 million-track sample.

The 45–180 day peak window was selected in consultation with RUN. The lower bound ensures that the model focuses on songs whose breakout occurs after the initial release period, while also allowing early streaming signals and the Luminate reporting lag to stabilise before a potential acquisition decision. The upper bound limits the analysis to early-viral trajectories whose pre-peak dynamics can plausibly be inferred from limited early observations. Songs peaking much later typically reflect delayed-break dynamics and are outside the scope of this thesis.

3.2 Viral Labelling Criteria

A song is labelled as viral if it satisfies all of the following criteria, applied directly within the SQL construction step:

- **Peak timing:** the all-time peak streaming day falls between day 45 and day 180 post-release.

- **Peak magnitude:** peak daily streams are at least 5,000.
- **Sustained popularity:** the song records at least 7 days with streams above 50% of peak volume.
- **Baseline activity:** median daily streams are at least 500 across the first 90 days.
- **Continuity:** there are no more than 3 days with zero streams after release.
- **Data completeness:** at least 98% of days between release and last observation are present.
- **History length:** the song has at least 230 days of observation if released between 2019 and 2024, and at least 90 days if released in 2025 or later.

These criteria are designed to identify songs with genuine post-release breakout dynamics rather than launch-day peaks, isolated one-day spikes, reporting artefacts, or catalogue anomalies. Applying them identifies 19,140 viral songs, which form the positive class used throughout the predictive modelling framework in Section 4.

3.3 Viral Trajectory Shapes

Although the labelling rule assigns a single viral label, the resulting songs do not all follow the same trajectory. Figure 1 illustrates this heterogeneity by grouping peak-normalised viral curves into representative trajectory shapes. Because each curve is scaled by its own peak, the comparison reflects shape rather than magnitude. Approximately 2,000 viral songs are excluded from this exploratory visualisation because they do not satisfy the 230-day minimum series length used for shape comparison; this exclusion affects only the figure, and the full viral population is retained for downstream modelling.

The figure shows that viral songs differ in rise speed, decay speed, persistence, and peak timing. Some songs rise sharply and decay quickly, while others build more gradually, sustain elevated streaming for longer, or resemble plateau-like patterns. Importantly, visible differences in growth, acceleration, and persistence already appear within the first months after release.

These patterns motivate five modelling choices in Section 4. First, the proposed pipeline reconstructs an explicit forward trajectory rather than predicting only aggregate post-purchase streams. Second, post-peak decay is estimated separately within peak-magnitude buckets rather than through a single global decay rate. Third, the visible early differences support using the first 35 days as the observation window. Fourth, because peak timing varies within the 45–180 day window, the valuation framework uses a representative median peak day as a simplifying assumption. Fifth, the pre-peak path from the observation window to the representative peak day is approximated by a linear ramp from the observed day-35 baseline to the predicted peak level.

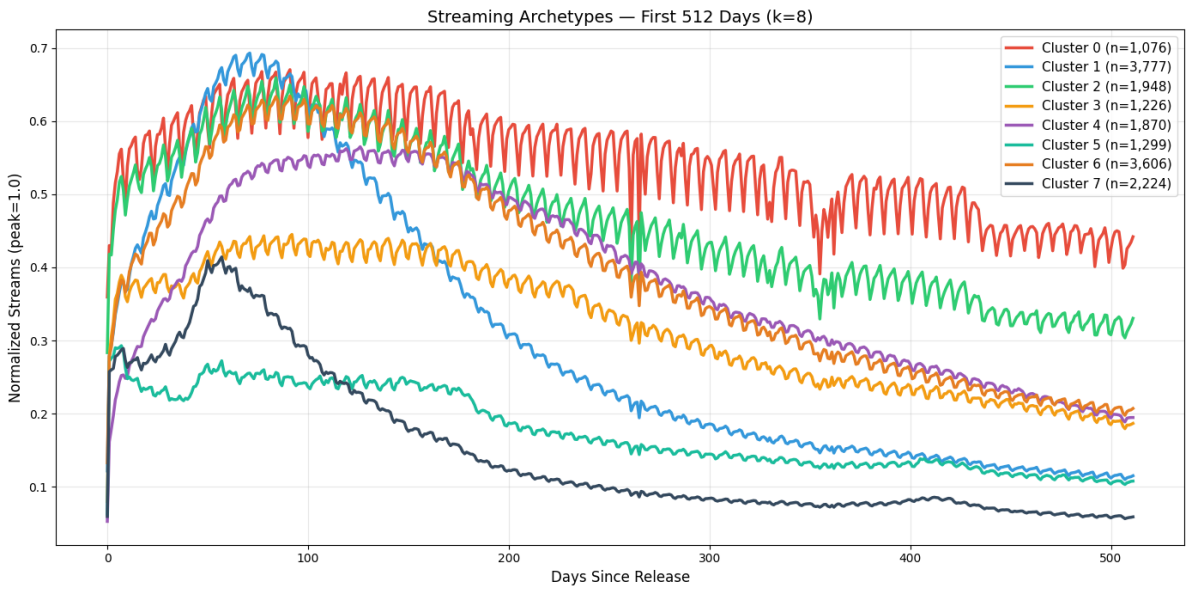

Figure 1: Representative viral trajectory clusters. Streaming trajectories are normalised by each song's own peak value.
Source: Luminate

3.4 Modelling Assumptions and Final Splits

The framework also makes two valuation simplifications. First, royalty rates vary across DSPs, geographies, and rights-share arrangements, but Section 4.5.2 uses a constant per-stream rate as a blended portfolio-level approximation. Second, acquisition cost is modelled through a naive flat-then-decay fair market price benchmark. This decouples the cost benchmark from the predictive model, so the model's economic value is measured by the gap between projected revenue and the naive benchmark price.

The final modelling dataset comprises 148,038 tracks, of which 19,140 are labelled viral. The dataset is split chronologically at 1 June 2024: 111,850 tracks released before this date form the training era, while 36,188 tracks released on or after this date form the hold-out test set. This chronological split prevents future information from entering model training and mirrors a realistic deployment setting.

The training era is further divided into training, validation, and calibration subsets using a 75/10/15 split. The validation set is used for early stopping, while the calibration set is used for isotonic probability recalibration and conformal prediction intervals, as described in Sections 4.3.3 and 4.4.3. Because the calibration set is drawn exclusively from the training era, it remains strictly prior to the test set, preserving the temporal ordering of training, calibration, and testing.

4 Methodology

4.1 Modelling Framework Overview

Let $y_{i,t}$ denote the daily streaming volume of song i on day t since release. The sequence $\{y_{i,t}\}_{t=1}^{T_i}$ defines song i's streaming trajectory.

The framework predicts future streaming performance from early consumption dynamics and translates these predictions into valuation-relevant quantities. It has two stages: a classifier estimating the probability of a viral trajectory, and a forecaster estimating the distribution of post-observation outcomes. Both operate on the same 35-day post-release window and feed into the valuation layer (Section 4.5); portfolio construction is treated downstream (Section 5).

Two forecasting formulations are compared. V1 predicts multiple quantiles of a peak-to-baseline ratio, from which a forward path is reconstructed and integrated to obtain revenue. V2 predicts total post-purchase stream volume directly as a point estimate. The comparison isolates the empirical value of trajectory modelling against end-to-end regression on the integrated outcome.

4.2 Data Splitting and Feature Construction

Following Section 3, the dataset is split chronologically into a training era and a hold-out test set, with the training era further divided into S_{tr}, S_{val}, and S_{cal}. Validation supports early stopping and model selection; calibration supports isotonic probability recalibration and split-conformal quantile calibration. Since S_{cal} is drawn from the training era, calibration remains chronologically prior to the test set.

All models use a fixed $w = 35$ day observation window—long enough for early trajectory differences to emerge, short enough for early acquisition decisions. Detectability is also evaluated at $w \in \{14, 21, 28\}$.

Two feature representations are derived. The classifier consumes a 10-dimensional representation built on running-peak normalisation (each daily count divided by the running maximum). From this, we construct features capturing growth, momentum, volatility, and acceleration: relative growth rates, 7- and 14-day rolling means and their ratio, rolling volatility, cumulative averages, second differences, a breakout indicator, and normalised time since the running peak. Running-peak rather than global-peak normalisation avoids forward-looking bias, as the global peak is typically unobserved at day 35.

The regressor extends this with four magnitude-sensitive features: $\log(1 + \text{streams})$, recent

velocity, the 7-day coefficient of variation, and a normalised local trend slope. These preserve scale information the classifier suppresses but the revenue forecast requires.

4.3 Classification Models

Two architectures are evaluated for virality classification: a custom bidirectional LSTM and a classifier built on the MOMENT-1-small foundation model.

4.3.1 LSTM Classifier

The LSTM classifier maps the $(35, 10)$ feature tensor to a single virality logit through a 128-dimensional input projection, a 4-layer bidirectional LSTM with hidden dimension 512 per direction and dropout 0.4, and a two-layer feedforward head over the concatenated final hidden states of the forward and backward passes. A schematic overview of the LSTM classifier is provided in Appendix [B].

Training minimises the positive-class-weighted binary cross-entropy

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{B}\sum_{i=1}^{B}\left[w_+\, y_i \log \sigma(\hat{z}_i) + (1 - y_i)\log\left(1 - \sigma(\hat{z}_i)\right)\right],$$

where $\sigma(\cdot)$ denotes the logistic sigmoid function, and $w_+ = n_-/n_+$ corrects for the approximately 13% viral base rate.

Optimisation uses AdamW (learning rate 10^{-3}, weight decay 10^{-4}) with cosine annealing over 50 epochs, gradient clipping at norm 1.0, mixed-precision computation, and early stopping on validation AUC with patience 10. Hyperparameters (hidden dimension, layers, dropout, projection and head dimensions, learning rate, weight decay, batch size) are selected via Bayesian optimisation with Optuna, using the Tree-structured Parzen Estimator (TPE) sampler and median pruning.

To improve robustness across early-detection horizons, batches are stochastically truncated during training: with probability 0.5, sequence length is sampled uniformly from $[14, 35]$. This regularises the model against overfitting to the fixed 35-day window and provides implicit supervision at shorter horizons.

4.3.2 MOMENT Foundation Model Classifier

For comparison, a classifier is built on the MOMENT-1-small foundation model, a pretrained encoder for general-purpose time-series representation. The input is a single-channel, peak-

normalised 35-day streaming sequence, zero-padded to the model's 512-token context with an attention mask. The pooled embedding is passed through the same feedforward head used in the LSTM model. A schematic summary of the MOMENT classifier pipeline is provided in Appendix [B].

Training proceeds in two stages. Stage 1 freezes the backbone and trains the head with AdamW (learning rate 10^{-3}), label smoothing 0.05, and cosine annealing over 30 epochs. Stage 2 unfreezes the backbone and fine-tunes the full model for 20 epochs with differential learning rates of 5×10^{-6} for the backbone and 10^{-4} for the head, preceded by a 3-epoch linear warmup. Gradient checkpointing fits training within the memory of a single NVIDIA T4 GPU. Effective batch sizes of 256 (stage 1) and 512 (stage 2) are achieved via gradient accumulation; the larger stage-2 batch stabilises fine-tuning once the backbone is unfrozen. Both stages use the same loss as the LSTM and select on validation AUC.

4.3.3 Probability Calibration

Neural network classifiers tend to produce miscalibrated probability estimates, particularly under class imbalance. Because the predicted virality probability is used downstream in expected-value computations and portfolio construction, calibration is necessary.

We apply isotonic regression (Zadrozny & Elkan, 2002) on the calibration set S_{cal}. Let $\hat{p}_i = \sigma(\hat{z}_i)$ denote the predicted probability for song i. Isotonic regression solves

$$\min_{\{\hat{y}_i\}} \sum_{i \in S_{\text{cal}}} (y_i - \hat{y}_i)^2 \quad \text{subject to} \quad \hat{y}_i \leq \hat{y}_j \text{ whenever } \hat{p}_i \leq \hat{p}_j,$$

yielding a non-decreasing step function $g : [0, 1] \to [0, 1]$ such that $g(\hat{p}_i) = \hat{y}_i$. The calibrated probability for any classifier output is then

$$p_i^{\text{cal}} = g(\hat{p}_i). \tag{1}$$

This effectively replaces raw network probabilities with empirically calibrated frequencies, improving the reliability of downstream expected-value computations. Because S_{cal} is strictly prior to the test set in time, no temporal leakage is introduced. The calibrated probabilities are used in all downstream portfolio decisions.

4.3.4 Classifier Interpretability

The primary research question asks not only which classifier produces more profitable acquisitions but also what temporal signals each classifier exploits, and how these compare across architectures. Answering the second half requires an attribution layer linking each classifier's output to its input. We apply three complementary post-hoc methods—input

gradients, integrated gradients (Sundararajan, Taly, & Yan, 2017) with a zero baseline, and non-overlapping 5-day temporal ablation—to both the LSTM ($x \in \mathbb{R}^{w\times 10}$) and MOMENT ($x \in \mathbb{R}^{w}$) classifiers. Attributions are computed on each model's calibrated top-50 candidates and aggregated separately for the five highest- and five lowest-profit acquisitions, isolating the signals each classifier uses to discriminate profitable from unprofitable picks rather than what it relies on across the candidate pool as a whole. Implementation details, model-specific ablation handling, and per-feature aggregations are deferred to Appendix [C].

4.4 Forecasting Models

A virality probability alone is insufficient for an acquisition decision: a song likely to break out can still be unprofitable if its predicted post-purchase streams are too small relative to cost. Two regressor formulations are therefore considered: the proposed trajectory-modelling regressor (V1) and a naive cumulative-sum baseline (V2). Each maps early-life dynamics to either a point estimate of cumulative streams or a set of conditional quantiles describing trajectory shapes.

Both models consume a $(w, 14)$ feature tensor from the first w days post-release and share an architecture; only the target, head, and loss differ. Easy negatives are excluded from training: at inference the regressor only sees candidates that have passed the classifier filter, so training on the same hard-candidate distribution aligns inputs with the deployment regime and avoids spending capacity on cases the classifier already rejects. The shared backbone is a bidirectional LSTM with a $14 \to 64$ input projection, 3 layers of hidden dimension 256, dropout 0.3, and an MLP head over the concatenated final hidden states. Training uses AdamW (learning rate 10^{-3}, weight decay 10^{-4}) with cosine annealing over 50 epochs, gradient clipping at norm 1.0, and mixed-precision computation. A separate regressor is trained at each $w \in \{14, 21, 28, 35\}$. A schematic is provided in Appendix [B].

4.4.1 Trajectory-Modelling Regressor (V1)

The proposed regressor predicts five quantiles of a forward-window peak-to-baseline ratio. With prediction horizon $H = 90$ and baseline floor $\varphi = 100$, the day-w baseline is

$$b_i = \max(\text{median}(y_{i,\,w-7:w})\,,\,\varphi)\,, \tag{2}$$

and the target is

$$r_i = \min\left(\frac{\max_{t\in[w,\,w+H]}\, y_{i,t}}{b_i},\; R_{\max}\right), \tag{3}$$

where $R_{\max} = 100$ caps extreme upper-tail ratios that would otherwise dominate the loss and destabilise training. This normalisation isolates relative growth dynamics from absolute scale, allowing the model to learn trajectory shape independently of baseline popularity.

Training minimises the multi-quantile pinball loss

$$\mathcal{L}_{\text{pinball}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{q \in \mathcal{Q}} \max(q(r_i - \hat{r}_{i,q}),\ (q-1)(r_i - \hat{r}_{i,q}))\,,$$

with $\mathcal{Q} = \{0.10, 0.25, 0.50, 0.75, 0.90\}$.

These forecasts form the basis of the valuation framework, where predicted trajectories are translated into revenue estimates and investment decisions.

4.4.2 Cumulative-Sum Baseline Regressor (V2)

The naive baseline replaces the quantile head with a single-output head and predicts the integrated post-purchase stream sum directly. With purchase day $d_p = 45$ and projection horizon $D_{\text{proj}} = 365$, the target is

$$u_i = \log\left(1 + \sum_{t=d_p+1}^{\min(T_i,\, D_{\text{proj}})} y_{i,t}\right), \tag{4}$$

where T_i is the song's observed history length. Songs with $T_i < w{+}10$ are excluded from training; songs with shorter histories than D_{proj} contribute partial sums and are right-censored, biasing the regressor toward shorter-horizon revenue. Unlike V1, V2 cannot impute the unobserved tail through a parametric decay model.

Post-purchase sums span roughly six orders of magnitude, so training in $\log(1 + \cdot)$ space keeps the loss scale-invariant and prevents large-revenue songs from dominating the gradient. Training uses the smooth L_1 (Huber) loss with $\delta = 1.0$, and predictions are back-transformed at inference as $U_i = \max(\exp(\hat{u}_i) - 1,\, 0)$.

V2 is a deliberate methodological reduction: V1's trajectory dynamics, decay modelling, and uncertainty quantification are all replaced by a single end-to-end mapping from early observations to aggregate outcomes.

4.4.3 Quantile Calibration

Raw quantile-regression outputs are not, in general, well calibrated: empirical coverage at the predicted q-th quantile may deviate from the nominal level q, even when the model captures the conditional shape correctly. Because the predicted quantiles are subsequently used in revenue forecasting and downstream expected-value computations, calibrated coverage is required.

We therefore apply split-conformal calibration (Romano, Patterson, & Candès, 2019) on the calibration set S_{cal}, computing a per-quantile additive adjustment from the empirical residual distribution. Let $\hat{r}_{i,q}$ denote the raw q-quantile prediction for calibration song $i \in S_{\text{cal}}$, and let r_i denote the realised peak-to-baseline ratio. The conformal residual $e_{i,q} = r_i - \hat{r}_{i,q}$ measures the per-quantile prediction error.

The per-quantile adjustment is given by the empirical q-quantile of these residuals with the standard finite-sample correction:

$$\Delta_q = \text{Quantile}_{\alpha(q)}(\{e_{i,q} : i \in S_{\text{cal}}\}), \quad \alpha(q) = \min\left(\frac{\lceil (n_{\text{cal}} + 1)\, q \rceil}{n_{\text{cal}}},\, 1\right), \tag{5}$$

where $n_{\text{cal}} = |S_{\text{cal}}|$.

The calibrated quantile prediction for a new song is then

$$\tilde{r}_{i,q} = \text{cummax}_q(\hat{r}_{i,q} + \Delta_q), \tag{6}$$

where the cumulative maximum is taken along the quantile axis to enforce $\tilde{r}_{i,0.10} \leq \tilde{r}_{i,0.25} \leq \tilde{r}_{i,0.50} \leq \tilde{r}_{i,0.75} \leq \tilde{r}_{i,0.90}$. This projection is necessary because independent per-quantile adjustments can produce crossing quantiles, violating the ordering required for valid prediction intervals and for the forward-curve construction used in the valuation stage.

The calibration is static: adjustments are fitted once on S_{cal} and held fixed at test time. Because S_{cal} is strictly prior to the test set in time, no temporal leakage is introduced. V2 produces no quantile structure and therefore receives no conformal calibration. This asymmetry—where V1 provides calibrated predictive intervals while V2 does not—is intrinsic to the modelling formulation rather than a design choice in the comparison.

4.5 Valuation Framework

The predictive layer of Sections 4.3 and 4.4 produces, for each song, a calibrated viral probability p_i^{cal} and either a vector of calibrated quantile predictions $\tilde{r}_{i,q}$ (V1) or a single point estimate U_i of the post-purchase stream sum (V2). The valuation framework translates these outputs into per-song expected revenue, acquisition cost, and profitability metrics. This is the layer at which model predictions become economically meaningful: each forecast is mapped into a dollar value comparable to a market price.

4.5.1 Decision Timing and Information Set

Investment decisions are made at the purchase day $d_p = 45$, ten days after the end of the $w = 35$-day observation window. The ten-day gap reflects an operational lag: the model

scores at day 35, the candidate enters a deal pipeline, and the transaction closes around day 45. Revenue accruing during this interval belongs to the seller and is therefore excluded from the buyer's projection.

All downstream valuation computations are restricted to the information set $\mathcal{F}_{i,w} = \sigma(y_{i,1}, \ldots, y_{i,w})$ generated by the streams observed up to day w, together with the calibrated viral probability p_i^{cal} and either the calibrated quantile predictions $\{\tilde{r}_{i,q}\}_{q\in\mathcal{Q}}$ (V1) or the back-transformed point estimate U_i (V2). The derived features consumed by the classifier and regressor (rolling means, growth rates, log-streams, etc.) are measurable functions of $\mathcal{F}_{i,w}$ and therefore introduce no additional information.

4.5.2 From Streams to Revenue

Streams are converted into revenue at a constant per-stream royalty rate $\rho = \$0.004$, reflecting the average post-aggregator payout per stream across major DSPs. For a song i with realised daily streams $y_{i,t}$, the actual revenue accrued from the purchase day to the projection horizon $D_{\text{proj}} = 365$ is

$$A_i = \rho \sum_{t=d_p+1}^{\min(T_i,\, D_{\text{proj}})} y_{i,t}, \tag{7}$$

where T_i is the song's observed history length. Songs with $T_i < D_{\text{proj}}$ contribute right-censored realised revenue: A_i reflects only the streams observed up to T_i, not the full holding period. The minimum-history filter applied for retrospective accuracy reporting (Section 4.5.7) bounds the censoring but does not eliminate it. Projected revenue follows the same mapping applied to a forecast stream trajectory: V1 integrates a reconstructed forward curve at each quantile (Section 4.5.3), while V2 obtains the integrated stream sum directly from the regressor (Section 4.5.5). The royalty rate ρ is identical across both pipelines, so all revenue figures — actual and projected, V1 and V2 — are denominated on a common scale.

4.5.3 Trajectory Reconstruction (V1)

V1 reconstructs an explicit forward streaming trajectory rather than predicting aggregate revenue directly. The calibrated quantile regressor estimates the conditional quantile function of the future peak-to-baseline ratio r_i given the early streaming history $\mathcal{F}_{i,w}$. The calibrated outputs approximate the inverse conditional CDF at the quantile levels $\mathcal{Q} = \{0.10, 0.25, 0.50, 0.75, 0.90\}$:

$$\tilde{r}_{i,q} \approx F^{-1}_{r_i|\mathcal{F}_{i,w}}(q), \qquad q \in \mathcal{Q}.$$

These five estimates provide a discrete summary of the conditional distribution of future peak intensity, ranging from downside to upside trajectory scenarios.

The calibrated quantiles are converted into forward streaming curves through peak reconstruction, ramp-then-decay assembly, and integration.

Peak reconstruction: Since r_i is defined as the ratio between the future peak and the day-w baseline b_i, the implied peak stream level at quantile q is $P_i^{(q)} = b_i \cdot \tilde{r}_{i,q}$. This anchors the predicted peak to the song's observed scale at the time of scoring.

Decay-rate estimation: Decay rates are estimated from training-era viral songs. Each song's post-peak tail — the streaming sequence from its global peak day onwards — is normalised by the peak value and truncated at 180 days. The pointwise median across these normalised tails defines a representative decay curve $m(\tau)$, where $\tau \in [0, 180]$ counts days since peak. An exponential rate is fitted by OLS on $\log m(\tau) \approx \alpha - \lambda\tau$, excluding days where $m(\tau) < 0.01$. The fit on the full pool of training viral songs yields the global decay rate λ_g, used in the FMP construction of Section 4.5.6. To let decay vary with peak intensity, viral songs are then stratified into six buckets via the five cut-points $\mathcal{Q} = \{0.10, 0.25, 0.50, 0.75, 0.90\}$ (the same five values used as the regressor's quantile levels in Section 4.4.1) on the training distribution of r_i, and the same procedure yields a bucket-specific λ_B within each bucket B. Buckets with fewer than 10 valid fitting points fall back to the median rate across successfully fitted buckets. The ramp-end time t_{ramp} is set to the empirical median of forward-window peak days across training viral songs, keeping it in the same temporal coordinates as r_i; the global peak is used only as a per-song normaliser for the decay tails.

Curve assembly: For each quantile q, the predicted ratio $\tilde{r}_{i,q}$ determines a decay bucket $B(q)$, and the forward streaming trajectory is constructed as

$$c_{i,t}^{(q)} = \begin{cases} y_{i,t}, & 1 \le t \le w, \\ b_i + (P_i^{(q)} - b_i)\left(\dfrac{t-w}{t_{\text{ramp}} - w}\right), & w < t < t_{\text{ramp}}, \\ P_i^{(q)} \exp\left[-\lambda_{B(q)}(t - t_{\text{ramp}})\right], & t_{\text{ramp}} \le t \le D_{\text{proj}}. \end{cases} \tag{8}$$

The curve therefore consists of observed history, a linear ramp to the predicted peak, and an exponential post-peak decay.

Per-quantile revenue: The revenue projection at quantile q is the integrated post-purchase curve scaled by the royalty rate:

$$R_{i,q} = \rho \sum_{t=d_p+1}^{D_{\text{proj}}} c_{i,t}^{(q)}. \tag{9}$$

The set $\{R_{i,q}\}_{q \in \mathcal{Q}}$ provides a discrete approximation to the conditional distribution of future post-purchase revenue.

4.5.4 Conservative Revenue and Expected Value (V1)

The five quantile-revenue projections are aggregated into a single operational summary. We define the conservative weighted revenue

$$\begin{aligned} R_i^{\text{cons}} = {} & 0.15\, R_{i,0.10} + 0.30\, R_{i,0.25} + 0.35\, R_{i,0.50} \\ & + 0.15\, R_{i,0.75} + 0.05\, R_{i,0.90}. \end{aligned} \tag{10}$$

This asymmetric weighting approximates a risk-averse decision rule, where downside scenarios receive greater emphasis than upside outcomes. In the context of fixed acquisition prices, overestimation leads directly to realised losses, whereas underestimation primarily results in missed opportunities. The specific weights are heuristic; sensitivity to the exact weighting is left to future work.

The expected revenue decomposes by the law of total expectation as

$$\begin{aligned} \mathbb{E}[R_i \mid \mathcal{F}_{i,w}] = {} & \mathbb{P}(\text{viral}_i \mid \mathcal{F}_{i,w})\, \mathbb{E}[R_i \mid \text{viral}_i, \mathcal{F}_{i,w}] \\ & + \mathbb{P}(\text{non-viral}_i \mid \mathcal{F}_{i,w})\, \mathbb{E}[R_i \mid \text{non-viral}_i, \mathcal{F}_{i,w}]. \end{aligned} \tag{11}$$

For the deterministic expected-value calculation, we approximate non-viral post-purchase revenue as negligible relative to acquisition cost, $\mathbb{E}[R_i \mid \text{non-viral}_i, \mathcal{F}_{i,w}] \approx 0$. This leaves $\mathbb{E}[R_i \mid \mathcal{F}_{i,w}] \approx \mathbb{P}(\text{viral}_i \mid \mathcal{F}_{i,w})\, \mathbb{E}[R_i \mid \text{viral}_i, \mathcal{F}_{i,w}]$. In practice, $\mathbb{P}(\text{viral}_i \mid \mathcal{F}_{i,w})$ is approximated by the calibrated classifier probability p_i^{cal}, while $\mathbb{E}[R_i \mid \text{viral}_i, \mathcal{F}_{i,w}]$ is approximated by the conservative quantile-weighted estimate R_i^{cons}. Hence,

$$\mathbb{E}[\text{Revenue}_i^{V1}] = p_i^{\text{cal}} R_i^{\text{cons}}. \tag{12}$$

This is a one-sided payoff approximation in which viral outcomes generate substantial revenue while non-viral outcomes contribute little relative to acquisition cost. Section 5.4 relaxes this approximation in the Monte Carlo validation by resampling non-viral realised revenues from the training-era empirical distribution.

4.5.5 Cumulative-Sum Baseline Revenue (V2)

V2 omits trajectory reconstruction entirely. Predicted revenue is obtained directly from the back-transformed regressor output, and the corresponding expected revenue applies the same calibrated probability used by V1:

$$R_i^{V2} = \rho \cdot U_i, \qquad \mathbb{E}[\text{Revenue}_i^{V2}] = p_i^{\text{cal}} R_i^{V2}. \tag{13}$$

Acquisition cost is computed by the same FMP model described below, so the comparison between V1 and V2 is symmetric on the cost side and isolates the contribution of trajectory modelling on the revenue side.

4.5.6 Acquisition Cost: Fair Market Price

The acquisition cost approximates the price a naive buyer would pay based on day-w streams alone, without conditioning on trajectory dynamics or breakout probability. The fair market price (FMP) is a stationary extrapolation: streams are held flat at the seed level $\bar{y}_{i,w}$ for a horizon $f = 30m$, then decay exponentially at a global rate λ_g. Operationally,

$$c_{i,t}^{\mathrm{FMP}}(m) = \begin{cases} \bar{y}_{i,w} & w+1 \le t \le w+f, \\ \bar{y}_{i,w} \cdot \exp(-\lambda_g(t-w-f)) & w+f < t \le D_{\mathrm{proj}}, \end{cases} \tag{14}$$

where $\bar{y}_{i,w} = \max(y_{i,\,w},\, \varphi)$ is the seed stream level (the day-w stream count), floored at φ to suppress vanishing-baseline pathologies. The FMP cost is the integrated curve scaled by the royalty rate:

$$\mathrm{FMP}_i(m) = \rho \sum_{t=w+1}^{D_{\mathrm{proj}}} c_{i,t}^{\mathrm{FMP}}(m). \tag{15}$$

The model is parameterised by the flat-extrapolation horizon $m \in \{1, 2, 3\}$ months and a premium multiplier $\mu \in \{1.00,\, 1.10,\, 1.15,\, 1.20,\, 1.25\}$ representing buyer-paid markups above the model FMP. The effective acquisition cost is

$$C_i = \mu \cdot \mathrm{FMP}_i(m), \tag{16}$$

with the operational (m, μ) cell selected at configuration time and sensitivity reported across all combinations. The FMP is intentionally decoupled from the model's revenue forecast: it represents the price floor implied by naive extrapolation, and the modelling pipeline's value is captured by the gap between FMP and trajectory-conditioned revenue.

4.5.7 Classifier Top-50 Comparison and Profitability Metrics

The valuation pipeline is evaluated on the test set by ranking songs in descending order of calibrated viral probability p_i^{cal} and retaining the top 50. This mirrors the intended screening use case: first identify songs most likely to break out, then assess whether the forecasting and valuation layer assigns economically meaningful revenue estimates to those candidates.

Per-song profitability is reported using return-on-investment (ROI) and annualised internal rate of return (IRR), with holding period $\Delta d = D_{\mathrm{proj}} - d_p = 320$ days and annualisation factor $365/\Delta d \approx 1.14$. Realised profitability uses actual post-purchase revenue A_i:

$$\mathrm{ROI}_i = \frac{A_i - C_i}{C_i}, \qquad \mathrm{IRR}_i = \left(\frac{A_i}{C_i}\right)^{365/\Delta d} - 1. \tag{17}$$

For V1, two projected variants substitute A_i above: the conservative quantile-weighted revenue R_i^{cons}, giving $\widehat{\mathrm{ROI}}_i^{\mathrm{cons}}$ and $\widehat{\mathrm{IRR}}_i^{\mathrm{cons}}$ (return conditional on a viral trajectory); and the

probability-weighted expected revenue $\mathbb{E}[\text{Revenue}_i^{V1}] = p_i^{\text{cal}} R_i^{\text{cons}}$, giving $\widehat{\text{ROI}}_i^{\text{EV}}$ and $\widehat{\text{IRR}}_i^{\text{EV}}$. The conditional metric captures upside magnitude given breakout; the expected-value metric combines magnitude with calibrated breakout probability. For V2, only the expected-value metrics are computed, with $\mathbb{E}[\text{Revenue}_i^{V2}] = p_i^{\text{cal}} R_i^{V2}$, since V2 produces no quantile structure and hence no analogue of R_i^{cons}.

IRR values are clipped to $[-1.0, 5.0]$ to suppress numerical pathologies from very small acquisition costs. Both ROI and IRR treat A_i as a single terminal payoff at D_{proj} rather than a stream of daily royalty inflows; the bullet-payoff approximation is consistent across V1, V2, and the FMP cost basis, preserving comparability. For retrospective valuation accuracy, the top-50 comparison is restricted to songs with at least 135 days of observed history (i.e. at least 90 days of post-purchase actuals); the filter is applied symmetrically to V1 and V2. The resulting per-song valuation outputs feed both the classifier top-50 comparison and the backtested investment strategies in Section 5.

5 Application: Walk-Forward Backtest

The valuation framework of Section 4.5 produces, for each test song, a calibrated viral probability, projected revenue, acquisition cost, and profitability metrics. There, these outputs feed a classifier top-50 comparison; this section instead operationalises them as budget-constrained walk-forward investment strategies under realistic execution conditions: candidates arrive sequentially in calendar time, capital is finite, and concentration must be managed across magnitude classes.

The backtest is a single-period proof of concept executed over the full test era (songs released on or after 1 June 2024). Two strategies are evaluated — a heuristic ranking strategy and a Kelly-inspired scoring rule — with V1 and V2 valuations feeding the same engine in parallel, and the resulting portfolios stress-tested through a Monte Carlo bootstrap. The purpose is to test whether the predictive and valuation layers can be operationalised as a deployable strategy under realistic constraints; multi-period validation across rolling retraining windows is left for downstream production deployment.

5.1 Walk-Forward Engine

The backtest processes test songs chronologically by scoring date

$$t_i^{\text{score}} = \tau_i^{\text{release}} + w, \tag{18}$$

where τ_i^{release} is the release date and $w = 35$ is the observation window. The predictive models are trained once on the pre-test training era and then deployed forward over the post-1 June 2024 test era.

At each scoring date, the predictive layer produces the calibrated probability p_i^{cal}, the relevant revenue forecast (V1 quantiles or V2 point estimate), the FMP-derived acquisition cost C_i, and the projected profitability metrics defined in Section 4.5. A candidate is admitted to the investable universe if it satisfies five filters:

1. day-w streams fall within the 25th-to-95th-percentile range of training-era viral songs;
2. $p_i^{\text{cal}} \geq 0.65$;
3. projected revenue is at least \$5,000, using R_i^{cons} for V1 and R_i^{V2} for V2;
4. acquisition cost is no more than 20% of total budget;
5. release date information is present.

For retrospective evaluation, candidates are additionally required to have at least $d_p + 90$ days of observed history, so realised post-purchase revenue can be compared to projected

revenue. This requirement ensures meaningful realised-performance measurement and is not part of the information set available at the scoring date.

A candidate is purchased if its strategy score exceeds the strategy-specific threshold, its cost does not exceed remaining budget, the relevant magnitude bucket has not reached the diversification cap $K = 10$, and adding the candidate does not reduce running projected portfolio IRR — the cumulative expected revenue against cumulative acquisition cost across the running portfolio — below 30%. The portfolio IRR floor is forward-looking: a candidate with positive expected ROI may still be rejected if its inclusion lowers cumulative projected portfolio IRR below the target. Songs are indivisible, so each accepted candidate is bought at full cost C_i. The budget is committed once and not replenished by accruing revenue. C_i excludes transaction costs (legal, escrow, due-diligence), which apply uniformly across V1, V2, and FMP and so do not affect relative comparisons. Unless otherwise stated, the default backtest uses a total budget of \$500K, FMP horizon $m = 1$ month, and premium multiplier $\mu = 1.00$.

5.2 Heuristic Strategy

The heuristic strategy ranks candidates by projected expected ROI:

$$s_i^{\text{heur}} = \frac{\mathbb{E}[\text{Revenue}_i] - C_i}{C_i}. \tag{19}$$

For V1, $\mathbb{E}[\text{Revenue}_i] = p_i^{\text{cal}} R_i^{\text{cons}}$; for V2, $\mathbb{E}[\text{Revenue}_i] = p_i^{\text{cal}} R_i^{V2}$.

Candidates are sorted in descending s_i^{heur} and bought greedily subject to the engine's constraints: budget, diversification cap, and portfolio IRR floor. The minimum acceptable score is zero, so the strategy does not voluntarily acquire songs with negative projected expected ROI. Position size is implicit: each accepted candidate is bought at full cost C_i, so capital allocation is determined by the rank ordering and cost distribution. Unlike the Kelly strategy below, diversification and concentration are enforced solely through the engine's hard constraints, with no soft penalties in the score itself.

5.3 Kelly-Inspired Strategy

The second strategy uses a discrete fractional-Kelly score with additional portfolio-level penalties. The classical Kelly criterion (Kelly Jr., 1956; Thorp, 1969) for asymmetric payoffs can be written as

$$f^{\text{Kelly}} = \frac{pb - (1-p)a}{ba}, \tag{20}$$

where p is the win probability, b is the net gain ratio if the position succeeds, and a is the loss ratio if it fails.

For V1, the win ratio is based on conservative projected revenue, while the loss ratio is based on downside revenue at the 10th quantile:

$$b_i = \max\left(\frac{R_i^{\text{cons}}}{C_i} - 1,\ 0\right), \qquad a_i = \max\left(\frac{C_i - R_{i,0.10}}{C_i},\ 0.01\right). \tag{21}$$

The 1% floor on a_i caps the Kelly fraction when downside revenue exceeds acquisition cost; otherwise $a_i \to 0$ would drive the position size unbounded. The raw Kelly fraction is then

$$f_i^{\text{Kelly}} = \max\left(\frac{p_i^{\text{cal}} b_i - (1 - p_i^{\text{cal}}) a_i}{b_i a_i},\ 0\right). \tag{22}$$

The raw fraction is adjusted by four multiplicative factors that account for model uncertainty, concentration, and budget pressure:

$$\begin{aligned} \text{conf}_i &= \frac{1}{1 + (R_{i,0.75} - R_{i,0.25})/R_{i,0.50}}, \qquad \text{asym}_i = 1 + 0.2 \cdot \max\left(\frac{R_{i,0.90} - R_{i,0.50}}{R_{i,0.50} - R_{i,0.10}} - 1,\ 0\right), \\ \text{bucketpen}_i &= \frac{1}{1 + n_b}, \qquad \text{costpen}_i = \max\left(1 - \frac{C_i}{M_d},\ 0\right). \end{aligned} \tag{23}$$

Wider interquartile intervals reduce the score (`conf`), upside-skewed quantile fans receive a bonus (`asym`), and bucket and cost penalties enforce diversification across magnitude classes and discipline against single-song budget concentration (`bucketpen`, `costpen`). Here n_b is the number of songs already held in the candidate's magnitude bucket and M_d is the remaining budget at scoring date d.

The final Kelly-inspired score is

$$s_i^{\text{Kelly}} = \kappa\, f_i^{\text{Kelly}} \cdot \text{conf}_i \cdot \text{asym}_i \cdot \text{bucketpen}_i \cdot \text{costpen}_i, \tag{24}$$

with fractional-Kelly safety factor $\kappa = 0.5$ (MacLean, Ziemba, & Blazenko, 1992; MacLean, Thorp, & Ziemba, 2011). Candidates are eligible for purchase if $s_i^{\text{Kelly}} \geq 0.05$.

For V2, the point-estimate baseline produces no downside, upside, or confidence information. The strategy therefore sets $a_i = 1$, $\text{conf}_i = \text{asym}_i = 1$, and retains only the bucket and cost penalties.

5.4 Monte Carlo Validation

The walk-forward backtest produces one realised portfolio on the observed test set. To characterise the model-implied distribution of portfolio outcomes, the selected portfolio is evaluated through a Monte Carlo with $M = 10{,}000$ simulations, combining a parametric Bernoulli model for viral-event activation with a non-parametric bootstrap conditional on

activation: the former carries the viral-probability calibration, the latter avoids imposing a parametric form on the heavy-tailed revenue distribution. The simulation is deliberately a parsimonious risk model: cross-song correlation in viral events, time-varying market conditions, and royalty heterogeneity are not modelled. Activations are independent across songs, revenue draws within each empirical pool are taken with replacement, and $M = 10{,}000$ provides approximately 1,000 outcomes in the worst-10% tail — sufficient for $\mathrm{CVaR}_{0.10}$ to stabilise.

For each simulation and each song, the procedure draws a Bernoulli activation with probability equal to the song's calibrated viral probability, then samples a revenue from a stratified pool of training-era songs that share the candidate's pre-purchase characteristics (peak-to-baseline ratio and day-w stream level for V1, predicted-revenue magnitude for V2). The simulated revenue is therefore the actual realised revenue of a comparable training-era song rather than a parametric draw. Summing across the portfolio yields one simulated outcome; repeating M times produces the distribution against which the deterministic backtest is compared.

Formally, let $\mathcal{P}$ denote the portfolio. For each $m = 1, \ldots, M$ and $i \in \mathcal{P}$, $Z_{i,m} \sim \mathrm{Bernoulli}(p_i^{\mathrm{cal}})$ determines whether song i realises a viral-like payoff, with revenue drawn from a viral or non-viral empirical pool constructed from training-era outcomes:

$$\tilde{A}_{i,m} = Z_{i,m} A_{i,m}^{*,\text{viral}} + (1 - Z_{i,m}) A_{i,m}^{*,\text{non-viral}}, \qquad A_{i,m}^{*,\bullet} \sim \widehat{F}_{\mathcal{A}^\bullet(i)}. \tag{25}$$

If the joint stratification cell contains fewer than five observations, the pool falls back progressively to bucket-only, day-w-only, and finally a global pool.

To capture systematic downside not exposed by song-level resampling — coordinated demand shocks, platform disruptions, or macroeconomic conditions affecting all holdings simultaneously — each simulation includes a stylised portfolio-level shock:

$$S_m = \begin{cases} 1 - 0.30 u_m, & \text{with probability } 0.05, \\ 1, & \text{with probability } 0.95, \end{cases} \qquad u_m \sim \mathrm{Uniform}(0.5, 1.0). \tag{26}$$

The 5% probability and up-to-30% magnitude are stylised rather than calibrated, since the music-rights asset class lacks long-run drawdown statistics from which to fit such parameters; robustness to alternative settings is left to deployment-stage analysis.

Simulated portfolio revenue, profit, and total cost are then

$$\mathrm{Rev}_m^{\mathrm{port}} = S_m \sum_{i \in \mathcal{P}} \tilde{A}_{i,m}, \qquad \Pi_m = \mathrm{Rev}_m^{\mathrm{port}} - C^{\mathrm{port}}, \qquad C^{\mathrm{port}} = \sum_{i \in \mathcal{P}} C_i. \tag{27}$$

We report the percentile of the realised portfolio outcome within the simulated distribution and $\mathrm{CVaR}_{0.10}$, the mean simulated profit in the worst 10% of simulations; ROI and annualised IRR follow standard definitions. The seed is fixed at 42. The Monte Carlo's informativeness depends on stratification quality: V1's strata are defined by characteristics observable at scoring time independently of the model, while V2's strata are defined by V2's own predicted revenue — a difference Section 6.3 examines for its risk-diagnostic consequences.

6 Results

This chapter follows the four questions stated in the introduction. First, we evaluate whether the representation choice affects early viral identification. Second, we compare valuation performance after candidate filtering. Third, we test whether Kelly-inspired sizing improves deployment relative to a simpler heuristic rule. Fourth, we compare the trajectory-based valuation head (V1) with the direct cumulative-sum baseline (V2).

Across all results, V1 denotes the trajectory-based valuation pipeline, while V2 denotes the direct baseline that predicts post-purchase stream volume as a single point estimate. Both pipelines use the same classifier probabilities, royalty conversion, and fair-market-price acquisition benchmark defined in Section 4.5.

6.1 Question 1: Classification

Question 1 evaluates whether early viral identification differs between the feature-engineered LSTM representation and the MOMENT foundation-model representation. Table 1 reports the corresponding classification metrics. While the two models achieve similar full-sample AUC at the 35-day horizon, their performance diverges in the hard subset and among the top-ranked candidates, which are the most relevant observations for an acquisition setting.

Table 1: Stage-1 classifier metrics. LSTM and MOMENT are nearly tied on full-sample AUC, while LSTM is stronger on the hard subset and in the top-ranked tail.

Metric	**LSTM**	**MOMENT**
Test AUC (All)	0.7692	0.7723
Test AUC (Hard)	0.6226	0.5860
AUC @ 14d	0.7024	0.6847
AUC @ 21d	0.7303	0.7201
AUC @ 28d	0.7517	0.7516
AUC @ 35d	0.7692	0.7723
Top-25 precision (All)	92.0%	68.0%
Top-50 precision (All)	78.0%	74.0%
Top-100 precision (All)	73.0%	75.0%
Top-200 precision (All)	70.5%	70.0%

MOMENT is marginally ahead on full-sample AUC at 35 days, but LSTM performs better on the hard subset and in the narrow top-ranked tail. LSTM leads on Top-25 precision, Top-50 precision, and early-window AUC at 14 and 21 days. Since the acquisition problem is a shortlist-generation problem rather than a full-sample ranking exercise, these tail metrics

are more economically relevant than headline AUC alone. The classification results therefore support using LSTM as the stronger screening model in the subsequent economic evaluation, while the next subsection examines whether the selected candidates are valued accurately once translated into projected revenue and acquisition returns.

6.2 Question 2: Valuation after screening

Question 2 evaluates whether the selected candidates are valued accurately once classifier scores are translated into projected post-purchase revenue. The fixed top-50 comparison is used as a diagnostic: each classifier selects its 50 highest-probability songs, after which V1 and V2 produce portfolio-level revenue estimates for the selected cohort. This comparison does not impose the full walk-forward investment constraints of Section 6.3; instead, it isolates whether the valuation heads assign plausible economic value to the candidates surfaced by the classifier.

Table 2: Fixed top-50 valuation diagnostic across classifier–valuation combinations. V1 uses conservative quantile-weighted revenue; V2 uses direct point-predicted revenue.

	Realised		**Projected**		**Bias**			
Pipeline	**Viral**	**Rev.**	**Proj.**	**PW Proj.**	**Proj.**	**PW Proj.**	**U/O**	**Prof.**
LSTM × V1	**40/50**	**$4.62M**	**$4.56M**	**$3.91M**	**−1.3%**	**−15.3%**	**23/27**	26/50
LSTM × V2	39/50	$4.28M	$5.92M	$5.11M	+38.4%	+19.5%	12/38	22/50
MOMENT × V1	**37/50**	**$2.54M**	**$2.46M**	**$2.00M**	**−3.1%**	**−21.0%**	**25/25**	**31/50**
MOMENT × V2	37/50	$2.37M	$3.32M	$2.73M	+40.0%	+15.1%	22/28	28/50

Notes: Viral = viral songs out of 50; Rev. = actual realised revenue; Proj. = projected revenue; PW Proj. = probability-weighted projected revenue; Bias columns compare projections to actual revenue; U/O = under-/over-projected holdings; Prof. = profitable holdings under FMP-1M 100% acquisition cost.

Table 2 shows two patterns. First, the LSTM-selected cohorts contain larger realised winners, producing higher total realised revenue than the MOMENT-selected cohorts under both valuation heads. This extends the classification result from Section 6.1: LSTM is not only stronger in the top-ranked tail by precision, but also surfaces a more valuable fixed top-50 candidate set.

Second, the trajectory-based V1 head is substantially better calibrated at the portfolio level. For both classifiers, V1 predicted revenue is close to actual revenue, with prediction bias of −1.3% for LSTM and −3.1% for MOMENT. By contrast, V2 overstates actual revenue by approximately 38–40% for both classifiers. The U/O counts show the same pattern: V1 is close to balanced, whereas V2 overestimates most LSTM-selected songs.

Table 3 evaluates the same fixed cohorts under alternative acquisition-cost assumptions. The projected–realised gap is consistently smaller for V1 than for V2. V1 projected ROI closely

tracks realised ROI across the grid, while V2 remains optimistic even when realised returns approach or fall below zero. This confirms that the trajectory-based valuation head is not only more accurate at the aggregate revenue level, but also more reliable once predictions are translated into acquisition returns.

The answer to Question 2 is therefore that V1 is the stronger valuation head. It produces less biased portfolio-level valuations and more robust realised ROI across acquisition-cost assumptions. The fixed top-50 diagnostic also reinforces the classification result: the LSTM-selected cohorts contain larger realised winners than the MOMENT-selected cohorts.

Table 3: Projected and realised ROI under selected acquisition-cost scenarios. Projected ROI uses conservative revenue for V1 and direct point-predicted revenue for V2.

Scenario	**Pipeline**	**Projected ROI**	**Actual ROI**
1M FMP, base price	LSTM × V1	+49.9%	+51.8%
	LSTM × V2	+97.6%	+42.8%
	MOMENT × V1	+44.1%	+48.7%
	MOMENT × V2	+88.9%	+34.9%
2M FMP, 15% premium	LSTM × V1	+14.9%	+16.4%
	LSTM × V2	+51.6%	+9.5%
	MOMENT × V1	+10.5%	+14.1%
	MOMENT × V2	+44.9%	+3.5%
2M FMP, 25% premium	LSTM × V1	+5.7%	+7.1%
	LSTM × V2	+39.4%	+0.8%
	MOMENT × V1	+1.7%	+4.9%
	MOMENT × V2	+33.3%	−4.8%
3M FMP, base price	LSTM × V1	+18.8%	+20.3%
	LSTM × V2	+56.7%	+13.2%
	MOMENT × V1	+14.2%	+17.9%
	MOMENT × V2	+49.8%	+7.0%
3M FMP, 15% premium	LSTM × V1	+3.3%	+4.6%
	LSTM × V2	+36.2%	−1.5%
	MOMENT × V1	−0.7%	+2.5%
	MOMENT × V2	+30.2%	−7.0%
3M FMP, 25% premium	LSTM × V1	−5.0%	−3.7%
	LSTM × V2	+25.3%	−9.4%
	MOMENT × V1	−8.6%	−5.7%
	MOMENT × V2	+19.8%	−14.4%

6.3 Question 3: Portfolio construction

Question 3 asks whether Kelly-inspired sizing improves deployment relative to simpler heuristic allocation. Table 4 reports the walk-forward backtest for the LSTM-screened candidate stream using a $500K budget, FMP-1M 100% acquisition cost, a per-song cost cap of 20% of budget, a per-bucket cap of 10 songs, and a 30% projected IRR floor. The classifier is fixed to LSTM since Sections 6.1 and 6.2 establish it as the stronger high-confidence screener; the

deployment exercise compares V1 against V2 under the two allocation rules.

Table 4: Walk-forward backtest results for the LSTM-screened candidate stream.

Head	Strategy	Songs	Cost	Actual Rev.	ROI	IRR	CVaR_{10}	MC pct.
V1	Heuristic	10	\$160K	\$242K	+50.8%	+60%	−\$22K	33rd
V1	Kelly	12	\$220K	\$899K	+307.9%	+397%	−\$47K	82nd
V2	Heuristic	18	\$499K	\$945K	+89.3%	+107%	+\$205K	66th
V2	Kelly	15	\$365K	\$599K	+64.0%	+76%	+\$151K	21st

Notes: CVaR_{10} is the mean simulated portfolio profit in the worst 10% of Monte Carlo outcomes. MC pct. is the percentile of the realised portfolio outcome within the model-implied simulated distribution.

For V1, Kelly-inspired sizing substantially improves realised performance: ROI rises from +50.8% to +307.9% and IRR from +60% to +397%. The Kelly score is only as informative as the uncertainty structure it consumes: V1's quantile head provides the dispersion and asymmetry signals the Kelly mechanism is designed to exploit, while V2's point-estimate head leaves the score with no distributional information to weigh. For V2, the pattern reverses: the heuristic rule outperforms Kelly (+89.3% vs +64.0%).

The Monte Carlo diagnostics qualify these numbers but require an asymmetric reading. V1 + Kelly's realised outcome lies at the 82nd percentile of its simulated distribution, indicating a right-tail draw under the model. V2 + Heuristic sits at the 66th percentile, closer to its model-implied centre. However, V2's MC is structurally over-optimistic: with no quantile head, V2's bootstrap pools are stratified by predicted revenue magnitude alone, and Section 6.2 shows V2 over-predicts by approximately 40%. This bias propagates into the MC, narrowing the simulated distribution and pushing even the worst-tail simulations into positive territory (positive CVaR). V1's CVaR is negative under both strategies because its quantile-based pools and dispersion-aware Kelly score carry the downside information V2 lacks. The V2 percentile and CVaR figures should therefore not be read as evidence of a safer portfolio — they reflect a tighter, upward-biased risk model rather than genuinely lower tail risk.

The answer to Question 3 is that Kelly-inspired sizing is useful when the valuation head provides meaningful uncertainty information. It improves deployment for V1, where quantile forecasts provide a basis for risk-adjusted ranking, but does not improve deployment for the point-estimate V2 baseline. The Monte Carlo comparison reinforces the same conclusion: V1's risk model produces the dispersion and downside signal that the Kelly mechanism is designed to exploit, while V2's risk model is too tight to be informative.

6.4 Question 4: Economic baseline

Question 3 has shown that V1's distributional outputs translate into stronger deployment outcomes; Question 4 examines whether this advantage holds at the underlying regression

accuracy level. This is the core comparison between V1 and V2. V1 represents the trajectory-based formulation: it predicts a distribution of future peak-to-baseline ratios, reconstructs a forward streaming path, and then converts the path into revenue. V2 represents the direct baseline: it predicts the post-purchase stream sum in one step.

Table 5: Regressor accuracy at the 35-day window.

Metric	**V1**	**V2**
MAPE	35.8%	96.0%
sMAPE	39.5%	57.9%
Prediction intervals	Yes	No

Notes: MAPE = mean absolute percentage error; sMAPE = symmetric mean absolute percentage error. Prediction intervals are produced by V1's quantile structure; V2 outputs only a point estimate. V1 is evaluated on the raw revenue scale after trajectory reconstruction; V2 is evaluated after back-transforming the log-space point prediction.

Table 5 shows that V1 is more accurate on the raw revenue scale. MAPE falls from 96.0% under V2 to 35.8% under V1, while sMAPE falls from 57.9% to 39.5%. V1 also provides prediction intervals through its quantile structure, with empirical coverage of 50.8% for the nominal 50% interval and 81.6% for the nominal 80% interval at the 35-day window. V2 has no corresponding interval output.

Table 6: Summary of V1 versus V2 across valuation and deployment results.

Dimension	**Winner**	**Evidence**
Revenue accuracy	V1	Lower MAPE and sMAPE on the raw revenue scale.
Portfolio bias	V1	Top-50 prediction bias is near zero for both classifiers; V2 overstates value by about 40%.
Acquisition returns	V1	Higher realised ROI across all selected FMP scenarios in Section 6.2.
Kelly deployment	V1	Kelly improves realised IRR under V1, but not under V2.
Simplicity	V2	Single point forecast with no curve reconstruction or conformal calibration.

Table 6 summarises the economic trade-off. V1 is more accurate, better calibrated, and more useful once forecasts are translated into acquisition returns and portfolio rules. V2 is simpler and retains some ranking power, but it is systematically more optimistic in the fixed top-50 valuation diagnostic and provides no uncertainty structure for sizing.

The evidence therefore favours the trajectory-based V1 formulation over the direct V2 baseline. The reason is not only statistical accuracy. The path-based approach preserves information about trajectory shape, peak intensity, uncertainty, and decay, while the direct baseline

collapses the future into one aggregate number. For music-rights valuation, where returns are skewed and acquisition errors are costly, this additional structure is economically useful.

6.5 Why LSTM and MOMENT pick different songs

The preceding sections establish that LSTM and MOMENT achieve nearly identical full-sample AUC at 35 days but diverge in the top-ranked tail. This subsection asks why. We first show that the two classifiers pick largely disjoint candidate sets with comparable economic value, then use linear probes to identify the axis on which the representations differ.

6.5.1 Portfolio disagreement analysis

Across the union of each model's V1 top-50, only 12 songs are selected by both classifiers; 38 are MOMENT-exclusive and 38 are LSTM-exclusive. A model trained as a noisier proxy for the other would produce a high-overlap candidate set. The observed 14% intersection indicates the opposite: the two representations rank candidates on materially different signals.

Table 7: Portfolio outcomes by selection regime on the 88-song union of LSTM and MOMENT V1 top-50 picks.

Metric	MOMENT-only	Consensus	LSTM-only
Composition			
n songs	38	12	38
Viral (label = 1)	27 (71%)	10 (83%)	30 (79%)
Economics (USD)			
Total cost	1,083,978	620,467	2,420,632
Total actual revenue	1,875,322	659,915	3,956,045
Avg actual revenue / song	49,351	54,993	104,106
V1 valuation (conservative blend)			
Total predicted revenue	1,526,458	929,697	3,627,585
Projected ROI	+40.8%	+49.8%	+49.9%
Forecast gap	−18.6%	+40.9%	−8.3%
Realized			
Real ROI	**+73.0%**	**+6.4%**	**+63.4%**

Note. Subset to songs with ≥ 135 days of post-release history. V1 predicted revenue is the conservative-quantile blend $0.15\,Q_{10} + 0.30\,Q_{25} + 0.35\,Q_{50} + 0.15\,Q_{75} + 0.05\,Q_{90}$ of the conformal-calibrated regressor at obs day 35, summed per group. This is the quantity the V1 buyer-side valuation pipeline produces.

Bucket-level economics show three patterns. Both exclusive buckets deliver strong ROI (+73.0% MOMENT-only, +63.4% LSTM-only), confirming the models identify valuable

songs along different axes. The consensus bucket lags at +6.4%. Per-song cost also diverges—LSTM-only averages $104k against MOMENT-only at $49k—so the models disagree on both which songs are viral and which price tier to enter.

V1 valuation patterns differ across subsets. Exclusive picks undervalue (−18.6% MOMENT-only, −8.3% LSTM-only), leaving the buyer-side blend conservative against truth. The consensus bucket inverts this, over-predicting by +40.9%—the unusual case where a deliberately downward-skewed valuation still overshoots. One explanation could be consensus picks are disproportionately near their peak, so V1's peak-to-baseline forecast overshoots in projection and leaves limited upside in the holding window, also supporting the underperformance of this strategy. Notably, with $n = 12$, this is suggestive rather than conclusive.

This does not contradict Question 2. As can be seen in Table 3, LSTM still outperforms when included in the portfolio construction consistently.

6.5.2 What features have the models encoded?

During training both models, regardless of architecture, encode predictive features for classification. To understand how these representations diverge, we fit OLS linear probes (5-fold CV, R^2) from each model's day-35 embedding to the ten engineered features. A baseline probe on the raw 35-day zero-padded stream series scores near zero, confirming the targets are not trivially recoverable. High R^2 means a simple readout extracts the feature, so training preserved it; low R^2 means it was discarded or compressed into a nonlinear geometry only the classifier head can decode.

Table 8: Linear-probe R^2 at obs_day = 35 for LSTM hidden state and MOMENT embedding.

Target	**LSTM hidden**	**MOMENT**	**Δ (MOMENT − LSTM)**
MOMENT recovers more linearly			
roll_14	0.746	0.935	+0.189
roll_7	0.764	0.932	+0.168
cum_avg	0.698	0.921	+0.223
days_since_max	0.556	0.844	+0.288
momentum	0.321	0.738	+0.417
normed	0.764	0.884	+0.120
breakout	0.638	0.722	+0.084
volatility	0.352	0.547	+0.195
LSTM recovers more linearly			
growth	0.142	0.023	−0.119
accel	0.110	0.024	−0.086

Part of this is mechanical: the LSTM receives `growth` and `accel` as inputs at each timestep, so their residual recoverability from the hidden state shows they survived rather than what

was learned. MOMENT works only from the normalized stream sequence and must construct everything its embedding preserves—that it linearizes the eight summary targets cleanly while discarding both derivatives is the substantive finding. Its representation organizes around level and trend rather than local high-frequency variation. The two models are therefore not noisy estimators of one signal but readers of different statistics off the same sequence: MOMENT exposes lifecycle position (cumulative trajectory, rolling means, time-since-peak), the LSTM exposes recent motion (`growth`, `accel`). The pick disagreement in Section 6.5.1 is the downstream consequence of two different feature dictionaries.

6.5.3 Timestep analysis: which part of the timeseries is most important to each model?

In order to see which part of the timeseries was most impactful for model classification, we employ temporal ablation and integrated gradients. Temporal ablation removes different 5-day blocks of the day-35 input and measures the score shift; a large drop means the model was reading from that block. The two architectures give opposite profiles (Figures 2 and 3, bottom panels). The LSTM's score collapses only when the trailing five days are masked; earlier blocks barely matter. MOMENT spreads sensitivity across the window, drawing comparable weight from the early and final days and less from the middle.

Integrated gradients on the LSTM (Figure 3, top and middle) sharpen the picture from the input side. `breakout` dominates for winners and losers alike, followed by `normed`, the rolling means, `cum_avg`, and `momentum`; `volatility` and `days_since_max` sit near zero, with `growth` and `accel` ranked low. The features the LSTM down-weights at input are exactly those its hidden state encodes least cleanly—a level-and-rolling-mean detector anchored on `breakout`, with derivatives carrying weight only as a marginal signal in the loser tail.

This sharpens the consensus underperformance in Section 6.5.1. A song at or near peak at day 35 satisfies both signatures by construction—most of its trajectory has accumulated, and the trailing window is still its largest segment—so the 12-song overlap is biased toward mature songs with little holding-window upside left. MOMENT-exclusive picks carry launch and mid-window signal the LSTM does not encode as effectively, meaning LSTM-exclusive picks on average pick fewer songs accelerating earlier than 35 days. The +73.0% / +63.4% / +6.4% ROI ordering could be interpreted that at 35 days, MOMENT is the best at picking out earlier breakouts among the buckets, while the consensus bucket is the regime where the evidence overlaps precisely because the song is already mature.

Individual timestep analysis of 4 different songs can be found in Appendix [D].

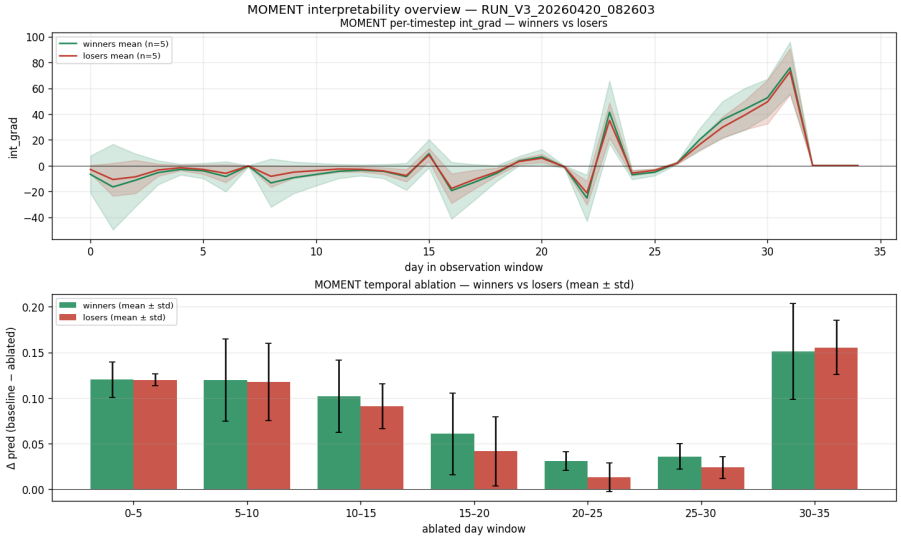

Figure 2: Classification MOMENT — per-timestep integrated gradients (top) and 5-day temporal-ablation Δpred (bottom) for n=5 winners and n=5 losers at obs day 35.

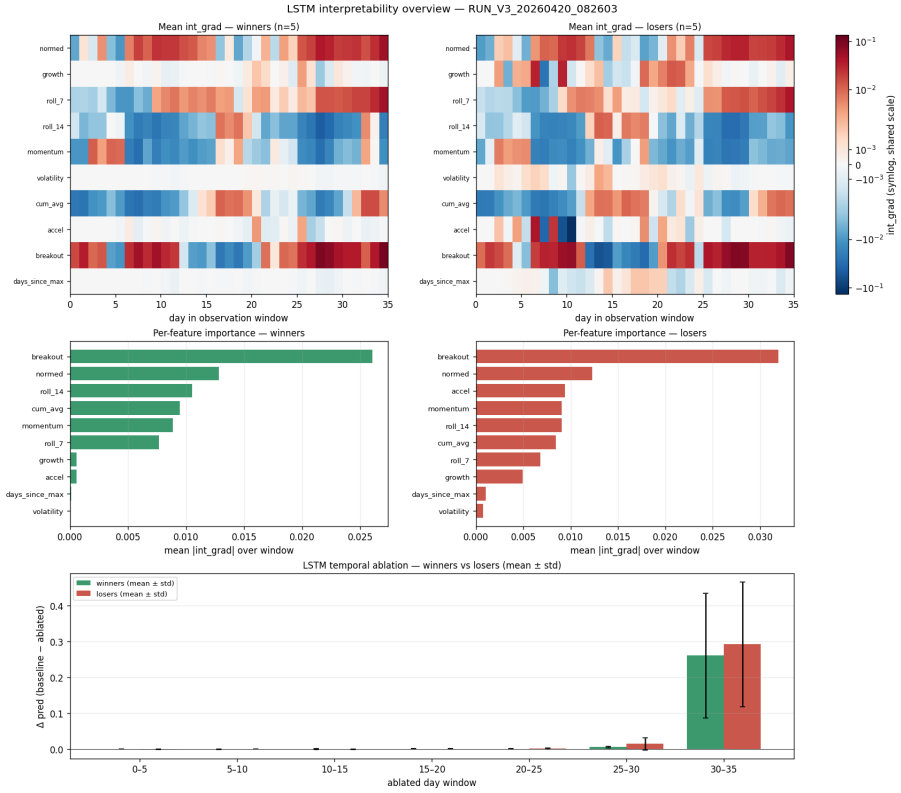

Figure 3: Classification LSTM — per-feature integrated gradients across the window (top), aggregated per-feature importance (middle), and 5-day temporal-ablation Δpred (bottom) for n=5 winners and n=5 losers at obs day 35.

7 Discussion

7.1 Interpretation of Mechanism

A consistent pattern runs through the four research questions: information about uncertainty matters more than information about central tendency, and the pipeline components that expose uncertainty downstream produce stronger economic outcomes than those that suppress it.

The classifier comparison illustrates the first part. MOMENT and the LSTM achieve nearly identical full-sample AUC at the 35-day horizon, but the LSTM dominates where it matters operationally—the hard subset, the early-detection windows at 14 and 21 days, and Top-25 precision. When the task is shortlist generation rather than full-sample ranking, the model that concentrates probability mass on the right candidates beats the one with marginally higher headline AUC. Feature engineering carries inductive biases toward the discriminative tail; MOMENT, lacking those biases at 35 days, distributes capacity more evenly across the population.

The probe analysis in Table 8 explains the pick disagreement that AUC obscures. Linear probes recover the engineered summary statistics—rolling means, cumulative averages, momentum, days-since-peak—substantially better from MOMENT's embedding than from the LSTM hidden state. The objectives explain this: masked reconstruction must preserve trajectory shape in linearly readable form, while end-to-end binary classification compresses the hidden state toward whatever subspace separates viral from non-viral. The LSTM additionally retains absolute stream values absent from MOMENT.

Temporal ablation tells the same story from the input side. MOMENT draws comparable weight from early and late days, consistent with summaries that require integrating across the full window. The LSTM only loses score when the trailing five days are masked, and its IG profile is dominated by `breakout` with slower-moving features near zero—an end-to-end classifier converged on recent-window evidence.

The two models therefore rank candidates on different axes drawn from different segments of the window; their top-50 picks are largely disjoint, and the representations are not substitutable. MOMENT's embedding carries more structurally readable information despite the LSTM being the stronger classifier in the operational tail.

The valuation comparison illustrates the second part. V1's trajectory-based head produces near-zero portfolio-level bias under both classifiers; V2's direct head over-forecasts under both. That the bias is invariant to classifier rules out the LSTM-versus-MOMENT axis as the explanation. The cause is structural: V2 minimises a loss on a heavy-tailed target, which pulls predictions toward the mean of the upper tail and produces persistent over-forecasting on songs with strong day-35 momentum. V1's quantile head is constrained at the

lower quantiles by the same conformal calibration that constrains the upper, anchoring the conservative weighting of the fan to realised downside.

The portfolio results carry the mechanism into deployment. V1 + Kelly substantially outperforms V1 + Heuristic, while V2 + Kelly underperforms V2 + Heuristic. Kelly's value depends on the dispersion and asymmetry signals it consumes; V1's quantile head supplies them, V2's point estimate does not. The same scoring rule that exploits useful uncertainty in V1 amplifies noise in V2. This is not a flaw of Kelly but a coherence requirement between valuation and sizing.

The Monte Carlo diagnostics confirm the reading from a different angle. V1's CVaR is negative under both strategies, signalling a risk model that retains downside information. V2's CVaR is positive—not evidence of a safer portfolio but of an upward-biased risk model. The bootstrap pools in V2's Monte Carlo are stratified by predicted revenue magnitude; since Section 6.2 shows that quantity is over-estimated, the bias propagates into the simulated distribution and pushes even worst-tail simulations into positive territory. The V2 Monte Carlo is internally consistent with V2's predictions but inherits their optimism.

7.2 Limitations

The pipeline rests on simplifying assumptions at multiple layers — the linear ramp to a homogenised median peak day, the bucketed exponential decay, the blended royalty rate, the binary viral label, the fixed 35-day observation window. These are design choices, not oversights. The thesis tests whether early streaming trajectories carry enough signal to support an end-to-end acquisition pipeline at all; that question is best answered by holding each layer at its simplest defensible specification and measuring whether the pipeline still produces economically meaningful outputs. Once that question is settled, refining each layer becomes a separable engineering problem rather than a precondition for the comparison. The limitations below are a roadmap of where further modelling would tighten dollar-level outputs, not flaws that undermine the comparative results.

The walk-forward backtest is single-period over the post-1-June-2024 test era — a proof of concept under realistic execution constraints, not a multi-window validation across rolling retraining periods. V1 + Kelly's realised outcome falls in the right tail of its own simulated distribution; the same strategy executed against many counterfactual realisations of the same statistical environment would on average produce returns substantially below the realised figure. The qualitative ordering across deployment metrics is robust — Kelly improves V1 outcomes, the heuristic does better with V2, V1 dominates V2 on calibration — but the magnitude of V1 + Kelly's outperformance is partly an artefact of favourable test-era realisations.

Three valuation approximations are inherited equally by both pipelines and therefore do not affect the V1-versus-V2 comparison, though they constrain absolute interpretation. Per-

stream royalties are blended at a single rate across DSPs, geographies, and rights structures, where real rates vary substantially. Post-peak decay is bucketed by peak magnitude but not by genre, release year, or artist tier. The 320-day holding window captures the early high-velocity phase of revenue but excludes the long catalogue tail that often dominates real-world deal returns. None of these threaten comparative validity, but they constrain how directly the reported ROI figures translate to a live acquisition.

Dataset size is itself a limitation. The training era contains roughly 112,000 songs, of which about 13% are viral — sufficient to learn the discriminative signal at the population level but tight for sharpening the conformal calibration on the V1 quantile head. A larger calibration set would shrink the finite-sample correction that currently widens predicted intervals and push empirical coverage closer to nominal at every quantile. A larger training set would benefit the regressor more than the classifier, since the regression target has heavier tails than the classification label and therefore requires more samples per region of the target distribution. The MOMENT comparison is particularly affected: foundation models scale favourably with data, and the present setup is well below saturation for either model. A substantially larger dataset, ideally with longer per-song histories so that observation windows beyond 35 days become available, is the single change most likely to improve calibration and forecast accuracy simultaneously.

A separate caveat concerns evaluation distribution. Both training and test sets are drawn from the same labelling procedure, which filters the broader Luminate universe through SQL heuristics on growth, peak timing, and median activity. The models are evaluated on the distribution of songs that procedure considers candidates, not on the distribution a real deployment would encounter. The reported precision and ROI figures should be read as upper bounds on what the same models would achieve against an unfiltered universe.

7.3 Implications for RUN

Four practical implications follow for RUN's investment workflow.

First, the V1 valuation head is materially closer to a deployable acquisition-decision input than V2. Across the FMP grid, V1's projected ROI tracks realised ROI closely; V2's projected ROI is consistently far above realised. For a buyer who must commit capital based on the model's price signal, the V2 forecast compresses margin of safety to the point where modest premium changes flip deal economics from positive to negative. The V1 conservative quantile-weighted figure (R_i^{cons} in Section 4.5.4) is the more defensible input for a reservation price.

Second, the system is best positioned as decision support rather than as an autonomous capital-allocation engine. The classifier surfaces a tractable shortlist; the valuation head produces a price signal that, in V1 form, can be benchmarked against an actual market quote; the Kelly sizing rule allocates within the shortlist conditional on the calibrated dispersion. None of these layers eliminates the need for human judgement on the surfaced candidates,

but each reduces the cost of that judgement by narrowing the search space and quantifying the uncertainty around the price.

Third, the disagreement analysis suggests a horizon-dependent allocation between architectures. At observation windows shorter than 35 days, the LSTM's hand-crafted features (acceleration, growth-linearity, peak-to-ignition ratios) explicitly encode what early-stage viral trajectories look like, whereas MOMENT must infer the same structure from a sparser input. For RUN's earliest scouting tier, where the cost of a missed breakout is the loss of the entire upside, the LSTM is the more reliable front-of-funnel screen.

Fourth, the logic reverses at longer horizons. Once several weeks of history are available, MOMENT's pretrained representations capture temporal patterns—persistence, regime changes, secondary peaks—that the engineered feature set was not designed for, and 60-day disagreement cases disproportionately favour it. The cost is interpretability: an LSTM flag is much easier to take to an investment committee and explain, while MOMENT requires much more interpretability work.

7.4 Future Work

Three methodological extensions follow directly from observations in the results.

V1's empirical coverage is close to nominal but slightly conservative. A monotone post-hoc calibration mapping from predicted to realised quantiles, fit on the calibration split, would tighten coverage at every level and reduce the conservatism penalty embedded in the current R_i^{cons} weighting. This is a natural next step before any production deployment.

The current backtest is single-period. Extending the framework to a rolling walk-forward design — where models are periodically retrained on expanding windows and the strategy is re-executed on each new release cohort — would replace the single right-tail observation with a distribution of realised outcomes across multiple test eras. The current ROI numbers would then be reportable with sample-based confidence intervals rather than as point estimates qualified by Monte Carlo percentile.

The 35-day observation window is constrained by the operational deal-closure timeline rather than chosen as a free hyperparameter. If the deployment context permitted longer observation horizons, applying MOMENT to the forecasting task — not just classification — becomes the relevant comparison. MOMENT was pre-trained on time-series forecasting, and the V1 quantile-regression task is a forecasting problem in everything but name. The thesis does not provide evidence on this comparison; whether MOMENT outperforms an LSTM-based regressor at longer horizons is an open empirical question that the current setup cannot answer.

8 Conclusion

This thesis asked whether a foundation-model representation of raw streaming trajectories outperforms a recurrent model trained on domain-engineered features in early-stage music rights valuation, and whether either approach can support an economically credible acquisition workflow from early streaming data. Using daily Luminate streaming histories for 148,038 tracks, split chronologically at 1 June 2024, the thesis addressed the problem in four stages: classification, valuation, portfolio construction, and economic baseline comparison. The conclusion proceeds in the same order, restating each research question and answering it directly, before stating what the combined evidence implies for the overarching research question.

Research Question 1: Does model choice affect viral candidate identification from the first 35 days of streaming data?

The answer is yes, but the direction depends on which part of the score distribution matters. At the 35-day observation horizon, MOMENT and the LSTM achieve nearly identical full-sample AUC (0.7723 versus 0.7692). However, the LSTM is materially stronger on the parts of the problem most relevant to an acquisition setting: the hard subset that early-stream baselines cannot separate (AUC 0.6226 versus 0.5860), early detection at 14 and 21 days, and Top-25 precision (92.0% versus 68.0%). Since acquisition is a shortlist-generation problem rather than a full-sample ranking exercise, the LSTM is the stronger screening model under these constraints.

Research Question 2: Does the LSTM-selected or MOMENT-selected top-50 songs produce stronger realised investment outcomes?

The answer is that the LSTM-selected cohort dominates on absolute return under both valuation heads, while MOMENT achieves a higher hit-rate at lower scale. LSTM realises \$4.62M and \$4.28M under V1 and V2 against \$2.54M and \$2.37M for MOMENT, with the ROI ordering preserved across every acquisition-cost scenario. MOMENT carries more profitable holdings overall, reflecting a flatter return distribution. The two classifiers therefore generate economically distinct portfolios: LSTM concentrates value in a small number of large winners, MOMENT spreads it across a broader profitable base. Where total realised revenue and capital-weighted return are the primary objectives, LSTM is the stronger screener regardless of valuation head.

Research Question 3: Do Kelly-inspired sizing rules turn these signals into stronger investment performance than simpler heuristic allocation rules?

The answer is conditional on the valuation layer. For V1, Kelly-inspired sizing improves realised performance substantially over the heuristic rule, lifting ROI from +50.8% to +307.9% and IRR from +60% to +397%. For V2, the pattern reverses: the heuristic rule outperforms Kelly (+89.3% versus +64.0%). The Kelly mechanism is only as informative as the uncer-

tainty structure it consumes. V1's quantile head supplies the dispersion and asymmetry signals the score is designed to exploit; V2's point-estimate head leaves the score with no distributional information to weigh. The Monte Carlo diagnostics support this reading: V1's CVaR is negative and informative under both strategies, while V2's apparently safer (positive) CVaR reflects an upward-biased risk model rather than genuinely lower tail risk. V1 + Kelly's realised outcome falling at the 82nd percentile of the simulated distribution indicates a right-tail draw rather than a typical realisation, but the qualitative ordering is consistent across deployment metrics: Kelly is useful when, and only when, the valuation head exposes meaningful uncertainty.

Research Question 4: Is it better to forecast revenue through peak and decay projection, or to predict total future streams directly?

The answer is the trajectory-based formulation. At the regression-accuracy level, V1 reduces MAPE from 96.0% to 35.8% and sMAPE from 57.9% to 39.5% on the raw revenue scale, and additionally produces calibrated prediction intervals (50.8% empirical coverage at the nominal 50% level, 81.6% at the nominal 80% level). V2 has no comparable interval output. The path-based approach preserves information about trajectory shape, peak intensity, uncertainty, and decay, while the direct baseline collapses the future into a single aggregate number. For music-rights valuation, where returns are skewed and acquisition errors are costly, the additional structure is economically useful.

Overarching research question: *In early-stage music streaming prediction, do foundation-model representations of raw streaming trajectories outperform a recurrent model built on domain-engineered features, and does any performance advantage translate into better portfolio investment outcomes?*

The evidence supports a conditional answer. On representation, the foundation model does not beat the feature-engineered baseline at the 35-day horizon. The LSTM wins on hard-subset discrimination, early detection, and top-of-list precision—the metrics that matter for screening—even where headline AUC is close. On economics, what determines investment credibility is the valuation layer, not the backbone. V1's calibrated distributional forecasts produce less biased valuations, more reliable realised returns across acquisition-cost scenarios, and the dispersion Kelly-style sizing needs. The claim that generalises beyond this thesis is that valuation design dominates backbone choice once predictions become prices and weights.

The representation-probe analysis qualifies this. The LSTM is the stronger single screener, but it and MOMENT are not redundant: their V1 top-50 selections intersect on only 14%, sit at materially different price points, and both exclusive subsets are profitable. The two architectures encode the same input differently, and a strategy drawing on both could plausibly outperform either alone. A foundation-model representation did not, in this setting, beat the feature-engineered baseline as a single screener. Whether the two combine usefully, or whether either strengthens under longer observation windows, larger pre-training corpora, or domain-adapted foundation models, is open.

References

Bass, F. M. (1969). A new product growth model for consumer durables. *Management Science*, *15*(5), 215–227. doi: 10.1287/mnsc.15.5.215

Cochrane, J. H. (2005). The risk and return of venture capital. *Journal of Financial Economics*, *75*(1), 3–52. doi: 10.1016/j.jfineco.2004.03.006

Cuntz, A., Kos, M., & Soriano Valdes, L. (2025). *The rise of music investment.* WIPO Economics Blog. Retrieved from https://www.wipo.int/en/web/economics/w/blogs/the-rise-of-music-investment

Damodaran, A. (2012). *Investment valuation: Tools and techniques for determining the value of any asset* (3rd ed.). Wiley.

Davies, K., Cole, H., & Turner, D. (2022, 10). *Understanding two decades of music catalog purchases* (Tech. Rep.). CNMlab. Retrieved from https://cnmlab.fr/en/short-wave/understanding-two-decades-of-music-catalog-purchases/

Dimolitsas, I., Kantarelis, S., & Fouka, A. (2023). *SpotHitPy: A study for ML-based song hit prediction using Spotify.* Retrieved from https://arxiv.org/abs/2301.07978

Dixit, A. K., & Pindyck, R. S. (1994). *Investment under uncertainty.* Princeton University Press. Retrieved from http://www.jstor.org/stable/j.ctt7sncv

Fjellström, C. (2022). *Long short-term memory neural network for financial time series.* Retrieved from https://arxiv.org/abs/2201.08218

Gabaix, X. (2009). Power laws in economics and finance. *Annual Review of Economics*, *1*, 255–294. doi: 10.1146/annurev.economics.050708.142940

Giantsidi, S., & Tarantola, C. (2025). Deep learning for financial forecasting: A review of recent trends. *International Review of Economics & Finance*, *104*, 104719. doi: 10.1016/j.iref.2025.104719

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., & Dubrawski, A. (2024). *MOMENT: A family of open time-series foundation models.* Retrieved from https://arxiv.org/abs/2402.03885

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. doi: 10.1162/neco.1997.9.8.1735

IFPI. (2024). *Global music report 2024: State of the industry* (Tech. Rep.). International Federation of the Phonographic Industry. Retrieved from https://ifpi.org/wp-content/uploads/2024/04/GMR_2024_State_of_the_Industry.pdf

Kelly Jr., J. L. (1956). A new interpretation of information rate. *Bell System Technical Journal*, *35*(4), 917–926. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1956.tb03809.x doi: 10.1002/j.1538-7305.1956.tb03809.x

Lim, B., Arik, S. Ö., Loeff, N., & Pfister, T. (2020). *Temporal fusion transformers for interpretable multi-horizon time series forecasting.* Retrieved from https://arxiv.org/abs/1912.09363

MacLean, L. C., Thorp, E. O., & Ziemba, W. T. (Eds.). (2011). *The kelly capital growth investment criterion: Theory and practice* (Vol. 3). World Scientific. Retrieved from https://www.worldscientific.com/worldscibooks/10.1142/7598 doi: 10.1142/7598

MacLean, L. C., Ziemba, W. T., & Blazenko, G. (1992). Growth versus security in dynamic investment analysis. *Management Science*, *38*(11), 1562–1585. doi: 10.1287/mnsc.38.11.1562

Middlebrook, K., & Sheik, K. (2019). *Song hit prediction: Predicting billboard hits using Spotify data.* Retrieved from https://arxiv.org/abs/1908.08609

Romano, Y., Patterson, E., & Candès, E. J. (2019). *Conformalized quantile regression.* Retrieved from https://arxiv.org/abs/1905.03222

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, *311*(5762), 854–856. doi: 10.1126/science.1121066

Soares Araujo, C. V., Pinheiro de Cristo, M. A., & Giusti, R. (2019). Predicting music popularity using music charts. In *2019 18th ieee international conference on machine learning and applications (icmla)* (pp. 859–864). IEEE. doi: 10.1109/ICMLA.2019.00149

Stoikov, S., & Kosyuk, I. (2022). *Valuation of music catalogs.* Retrieved from https://arxiv.org/abs/2209.09719

Stoikov, S., Singla, A., Cetin, U., & Cendra Villalobos, L. A. (2026). *Music as an asset class.* Retrieved from https://arxiv.org/abs/2602.05007

Sundararajan, M., Taly, A., & Yan, Q. (2017). *Axiomatic attribution for deep networks.* Retrieved from https://arxiv.org/abs/1703.01365

Thorp, E. O. (1969). Optimal gambling systems for favorable games. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, *37*(3), 273–293. Retrieved from http://www.jstor.org/stable/1402118

Universal Music Group. (2020, 12 7). *Universal music publishing group acquires Bob Dylan's entire catalog of songs.* Universal Music Group. Retrieved from https://www.universalmusic.com/universal-music-publishing-group-acquires-bob-dylans-entire-catalog-of-songs/

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth acm sigkdd international conference on knowledge discovery and data mining* (pp. 694–699). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/775047.775151

Zeng, Z., Kaur, R., Siddagangappa, S., Rahimi, S., Balch, T., & Veloso, M. (2023). *Financial time series forecasting using CNN and transformer.* Retrieved from https://arxiv.org/abs/2304.04912

Appendix A: Dataset Construction SQL Query

The following SQL query implements the consolidated assembly step described in Section 3.1:

```
CREATE OR REPLACE TABLE DATA_SCIENCE.THESIS_PROJECT.MODEL_DATASET AS
WITH release_peak AS (
    SELECT MR_ID,
           MIN(REPORT_DATE) AS RELEASE_DATE,
           MAX(REPORT_DATE) AS LAST_DATE,
           COUNT(*) AS track_days
    FROM DATA_SCIENCE.THESIS_PROJECT.SAMPLE_50M_TS_PEAK_45_180
    GROUP BY MR_ID
),
with_days_peak AS (
    SELECT s.MR_ID, s.DAILY_STREAMS,
           DATEDIFF(day, r.RELEASE_DATE, s.REPORT_DATE) AS
               days_since_release
    FROM DATA_SCIENCE.THESIS_PROJECT.SAMPLE_50M_TS_PEAK_45_180 s
    JOIN release_peak r ON s.MR_ID = r.MR_ID
),
qualified_nonviral AS (
    SELECT m.MR_ID
    FROM (
        SELECT MR_ID
        FROM with_days_peak
        WHERE days_since_release BETWEEN 0 AND 90
          AND MR_ID NOT IN (
              SELECT DISTINCT MR_ID
              FROM DATA_SCIENCE.THESIS_PROJECT.VIRAL_CANDIDATES
          )
        GROUP BY MR_ID
        HAVING MEDIAN(DAILY_STREAMS) >= 500
    ) m
    JOIN release_peak r ON m.MR_ID = r.MR_ID
    LEFT JOIN (
        SELECT MR_ID, COUNT(*) AS zero_days
        FROM with_days_peak
        WHERE DAILY_STREAMS = 0 AND days_since_release > 0
        GROUP BY MR_ID
    ) z ON m.MR_ID = z.MR_ID
    WHERE (z.MR_ID IS NULL OR z.zero_days <= 3)
      AND r.track_days >= DATEDIFF(day, r.RELEASE_DATE, r.LAST_DATE) *
          0.98
      AND ((YEAR(r.RELEASE_DATE) <= 2024 AND r.track_days >= 230)
           OR (YEAR(r.RELEASE_DATE) >= 2025 AND r.track_days >= 90))
),
release_broad AS (
    SELECT MR_ID,
           MIN(REPORT_DATE) AS RELEASE_DATE,
           MAX(REPORT_DATE) AS LAST_DATE,
           COUNT(*) AS track_days
    FROM DATA_SCIENCE.THESIS_PROJECT.SAMPLE_50M_TS
    GROUP BY MR_ID
```

```
),
with_days_broad AS (
    SELECT s.MR_ID, s.DAILY_STREAMS,
           DATEDIFF(day, r.RELEASE_DATE, s.REPORT_DATE) AS
              days_since_release
    FROM DATA_SCIENCE.THESIS_PROJECT.SAMPLE_50M_TS s
    JOIN release_broad r ON s.MR_ID = r.MR_ID
),
all_exclude AS (
    SELECT DISTINCT MR_ID
    FROM DATA_SCIENCE.THESIS_PROJECT.VIRAL_CANDIDATES
    UNION
    SELECT DISTINCT MR_ID FROM qualified_nonviral
),
easy_qualified AS (
    SELECT m.MR_ID
    FROM (
        SELECT MR_ID
        FROM with_days_broad
        WHERE days_since_release BETWEEN 0 AND 90
          AND MR_ID NOT IN (SELECT MR_ID FROM all_exclude)
        GROUP BY MR_ID
        HAVING MEDIAN(DAILY_STREAMS) >= 500
    ) m
    JOIN release_broad r ON m.MR_ID = r.MR_ID
    LEFT JOIN (
        SELECT MR_ID, COUNT(*) AS zero_days
        FROM with_days_broad
        WHERE DAILY_STREAMS = 0 AND days_since_release > 0
        GROUP BY MR_ID
    ) z ON m.MR_ID = z.MR_ID
    WHERE (z.MR_ID IS NULL OR z.zero_days <= 3)
      AND r.track_days >= DATEDIFF(day, r.RELEASE_DATE, r.LAST_DATE) *
          0.98
      AND ((YEAR(r.RELEASE_DATE) <= 2024 AND r.track_days >= 230)
           OR (YEAR(r.RELEASE_DATE) >= 2025 AND r.track_days >= 90))
),
easy_sampled AS (
    SELECT MR_ID, ROW_NUMBER() OVER (ORDER BY RANDOM()) AS rn
    FROM easy_qualified
)
SELECT *,
    CASE WHEN RELEASE_DATE < '2024-06-01' THEN 'train' ELSE 'test' END
        AS SPLIT
FROM (
    SELECT MR_ID, REPORT_DATE, RELEASE_DATE, DAYS_SINCE_RELEASE,
       DAILY_STREAMS,
           1 AS LABEL, 'viral' AS NEG_TYPE
    FROM DATA_SCIENCE.THESIS_PROJECT.VIRAL_CANDIDATES
    UNION ALL
    SELECT s.MR_ID, s.REPORT_DATE, r.RELEASE_DATE,
           DATEDIFF(day, r.RELEASE_DATE, s.REPORT_DATE) AS
              DAYS_SINCE_RELEASE,
           s.DAILY_STREAMS, 0 AS LABEL, 'hard' AS NEG_TYPE
```

```
    FROM DATA_SCIENCE.THESIS_PROJECT.SAMPLE_50M_TS_PEAK_45_180 s
    JOIN release_peak r ON s.MR_ID = r.MR_ID
    WHERE s.MR_ID IN (SELECT MR_ID FROM qualified_nonviral)
    UNION ALL
    SELECT s.MR_ID, s.REPORT_DATE, r.RELEASE_DATE,
           DATEDIFF(day, r.RELEASE_DATE, s.REPORT_DATE) AS
              DAYS_SINCE_RELEASE,
           s.DAILY_STREAMS, 0 AS LABEL, 'easy' AS NEG_TYPE
    FROM DATA_SCIENCE.THESIS_PROJECT.SAMPLE_50M_TS s
    JOIN release_broad r ON s.MR_ID = r.MR_ID
    WHERE s.MR_ID IN (SELECT MR_ID FROM easy_sampled WHERE rn <=
       100000)
)
ORDER BY MR_ID, DAYS_SINCE_RELEASE;
```

Appendix B: Model Architecture Figures

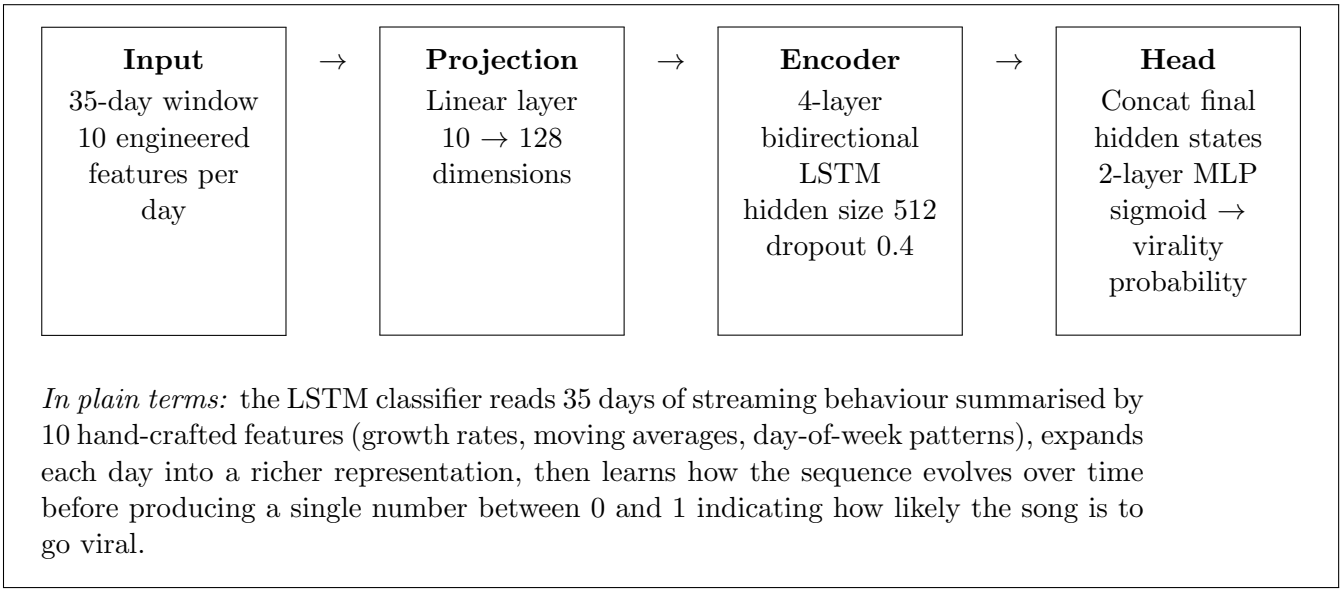

Figure 4: Architecture of the LSTM classifier used for early viral identification (Section 4.3.1).

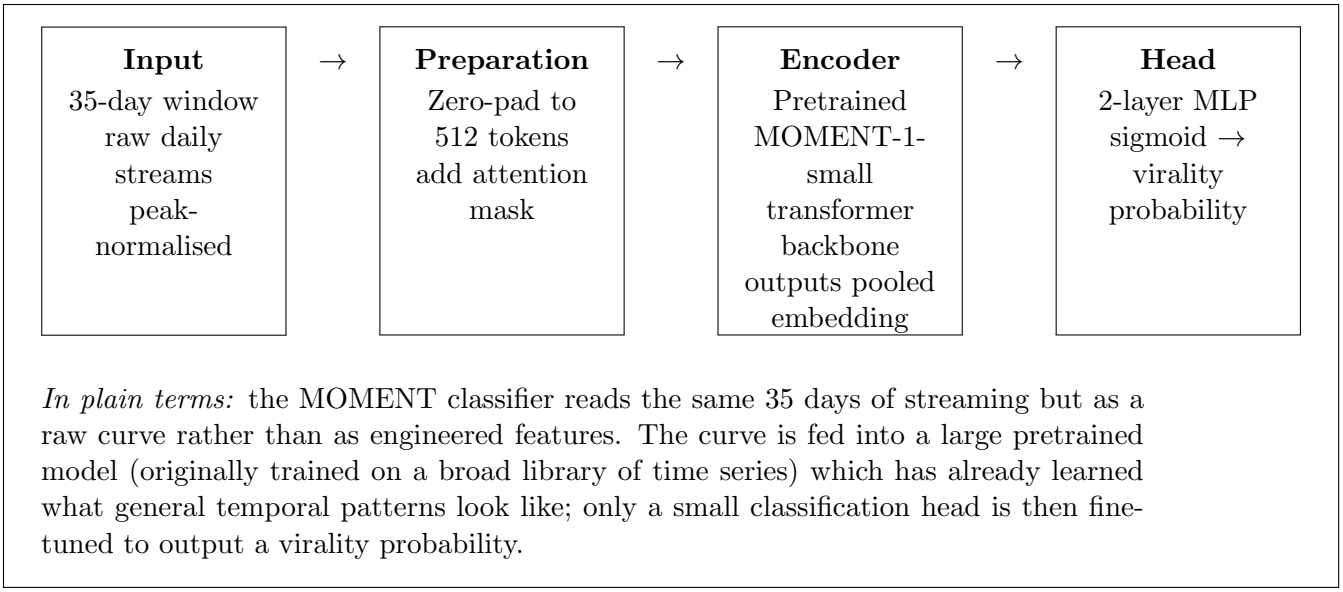

Figure 5: Architecture of the MOMENT classifier pipeline. Training uses a two-stage schedule (frozen-backbone warmup followed by joint fine-tuning); details in Section 4.3.2.

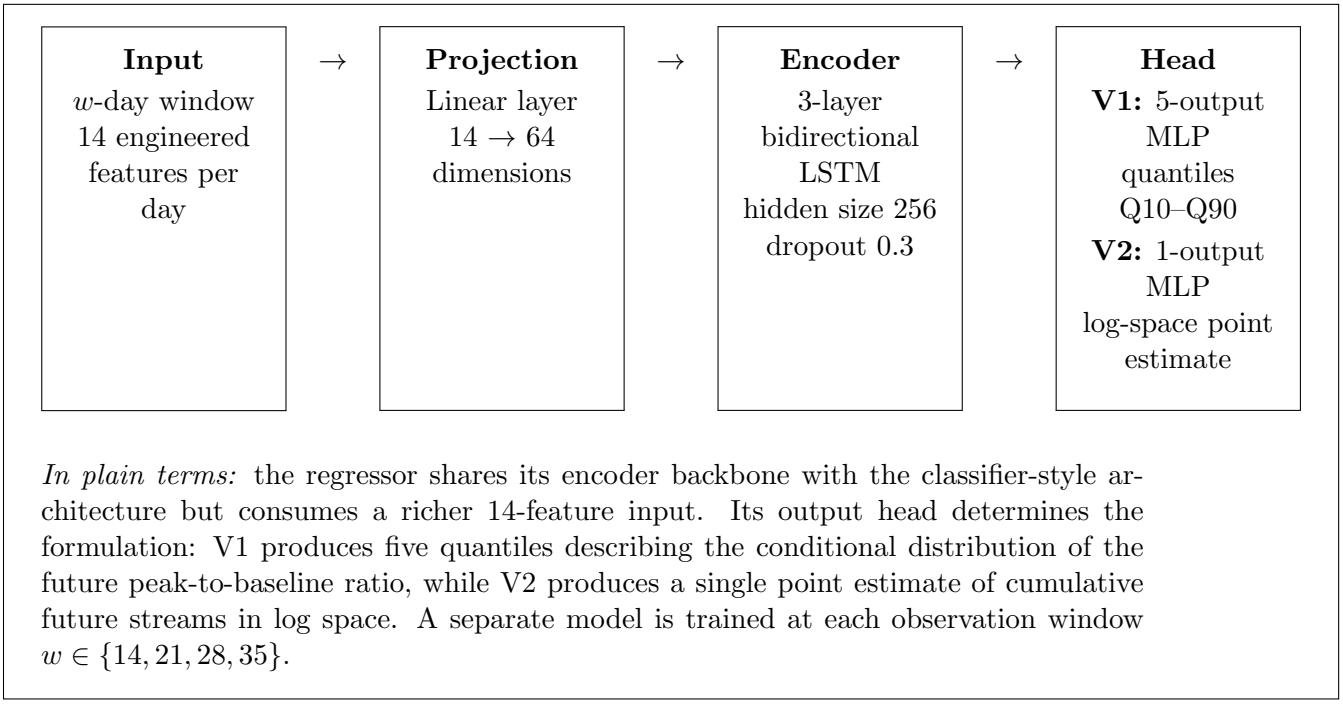

Figure 6: Architecture of the LSTM regressor used for both V1 (quantile peak-ratio prediction) and V2 (direct cumulative-stream point estimate). The shared encoder ensures the only difference between formulations is the output head and loss function (Sections 4.4).

Appendix C: Classifier Interpretability

We attribute the model output to its $w = 35$-day input via three methods, applied independently to the LSTM ($x \in \mathbb{R}^{w\times 10}$) and MOMENT ($x \in \mathbb{R}^{w}$) classifiers. Although the valuation and backtest pipelines use isotonic-calibrated classifier probabilities for expected-value calculations and portfolio construction, the attribution analysis is computed with respect to the pre-calibration model probability $\sigma(f(x))$. This avoids differentiating through the non-differentiable isotonic calibration map and keeps input gradients, integrated gradients, and temporal ablations on the same output scale.

Attribution Methods

Input gradient: The local sensitivity

$$g(x) = \frac{\partial \sigma(f(x))}{\partial x} \tag{28}$$

is computed by a single backward pass. Each entry of $g(x)$ measures the rate of change of the predicted probability with respect to a single input cell at the current point x. Large positive values indicate inputs that, if perturbed upward, would increase the predicted probability; large negative values indicate the opposite. Because $g(x)$ captures only the local slope at x, it can vanish even for inputs that are clearly important when the network is locally saturated or flat around the observed input, motivating the path-based extension below.

Integrated gradients: We use the zero baseline $x' = 0$ for both inputs: peak-normalised MOMENT streams at zero correspond to silence, while running-peak-normalised LSTM features have a no-signal baseline approximately at the origin of feature space. Integrated gradients accumulate local sensitivity along the straight-line path from baseline to input, so a feature can receive non-zero attribution if the model is sensitive to it anywhere along that path, not only at x itself. The line integral

$$\mathrm{IG}(x) = (x - x') \odot \int_0^1 \nabla_x \sigma(f(x' + \alpha(x - x')))\, d\alpha \tag{29}$$

is approximated via the trapezoidal rule with $m = 32$ subintervals (Sundararajan et al., 2017):

$$\mathrm{IG}(x) \approx (x - x') \odot \frac{1}{m} \sum_{k=0}^{m-1} \frac{1}{2} [\nabla_x \sigma(f(x_k)) + \nabla_x \sigma(f(x_{k+1}))]\,, \quad x_k = x' + \tfrac{k}{m}(x - x'). \tag{30}$$

The first factor $(x - x')$ scales each input cell by its displacement from the baseline; the second averages the model's sensitivity at $m + 1$ interpolation points along the path. Completeness,

$$\sum_{t,f} \mathrm{IG}(x)_{t,f} \approx \sigma(f(x)) - \sigma(f(x')), \tag{31}$$

holds up to discretisation error: the cell-level attributions add up to the total prediction shift between baseline and input, which makes individual cell magnitudes directly comparable within a song and on the same scale as the prediction itself. For MOMENT, the same completeness relation holds with the attribution sum taken over timesteps only.

Temporal ablation: For each non-overlapping window $W = [s, s+5)$ within the observation period, we record

$$\Delta_W = \sigma(f(x)) - \sigma(f(x^W)), \tag{32}$$

where x^W is the input with positions in W replaced by a fill value. Positive Δ_W means removing the information in W lowers the score, identifying W as a positive contributor to the original prediction; negative Δ_W flags windows whose removal increases the score, suggesting that these windows suppress the original prediction.

LSTM uses zero-fill ($x^W_{t,:} = 0$ for $t \in W$): features are pre-normalised so zero is a meaningful baseline. MOMENT cannot use zero-fill because RevIN normalisation re-centres the input on its per-instance mean and variance; zero-filling would distort both moments and contaminate the kept positions in the same forward pass. We therefore use mean-fill, replacing ablated positions with μ_{kept}, the mean of the kept positions. Mean-fill preserves the per-instance mean exactly and contributes zero to the variance sum-of-squares: after fill, the variance is

$$\frac{1}{w}\sum_{t \notin W}(x_t - \mu_{\text{kept}})^2 = \frac{w - |W|}{w}\sigma^2_{\text{kept}}. \tag{33}$$

Mean-fill therefore preserves the post-ablation mean exactly and only mechanically rescales the kept deviations by $\sqrt{w/(w-|W|)} = \sqrt{35/30} \approx 1.08$ relative to normalisation over the kept positions alone. This limits the artefact introduced by the ablation, whereas zero-fill would shift both the mean and variance and contaminate the retained timesteps through RevIN.

Pick selection and aggregation: From each model's top-50 candidate set used in the valuation analysis, the five highest- and five lowest-profit songs (by realised post-purchase profit $A_i - C_i$) define winners and losers. The contrast is diagnostic: comparing winners' attributions to losers' attributions helps diagnose which signals were associated with profitable high-confidence picks and which signals were associated with unprofitable high-confidence picks. For the overviews (Figures 2, 3), per-cell IG values are averaged within each role into mean-IG heatmaps on a shared SymLogNorm scale (linear near zero, log in the tails, used because IG distributions typically have a few dominant cells alongside many near-zero ones); per-feature mean |IG| is reported as importance bars summarising which features the LSTM relies on across timesteps; and 5-day ablation deltas are plotted as grouped bars (mean $\pm 1\sigma$ across the five picks per role).

Appendix D: Case Studies of song picks and interpretability

This appendix presents four representative attribution case studies drawn from the held-out test set, two for each architecture. Each figure stacks three panels: the full streaming trajectory with the 0–35 day observation window shaded and the day-45 purchase point marked; per-timestep integrated gradients attributions over the 35-day input; and a temporal ablation in which contiguous five-day windows are masked and the resulting drop in predicted revenue is plotted. Together they isolate where in the early window each model is actually drawing signal from, and how that allocation of attention coincides with downstream profitability.

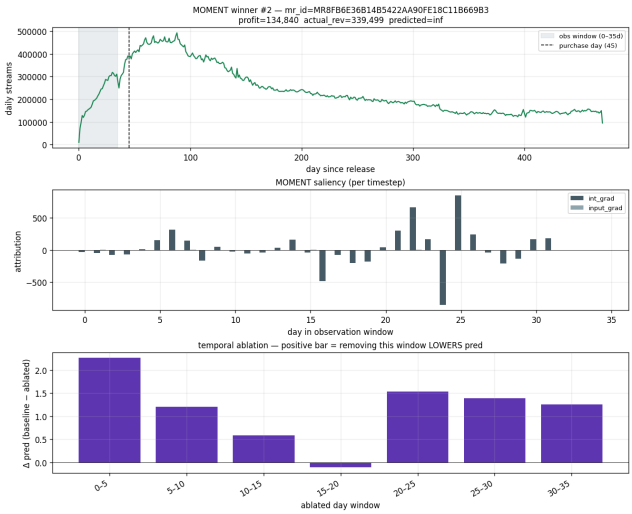

Figure 7: MOMENT, profitable pick. Δprofit = +134,840; realised revenue 339,499.

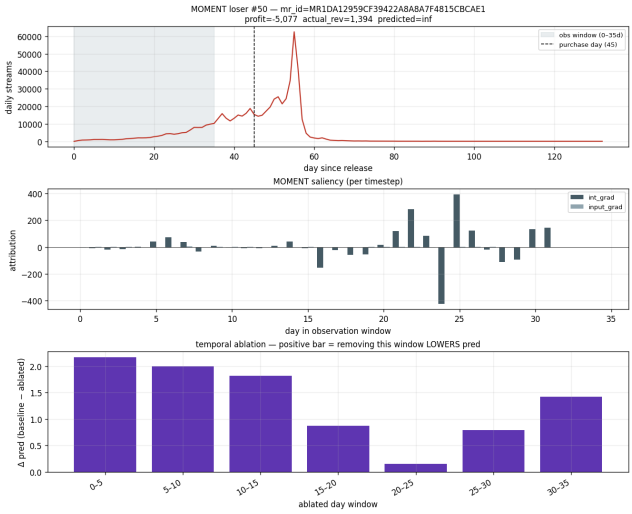

Figure 8: MOMENT, loss-making pick. Δprofit = −5,077; realised revenue 1,394.

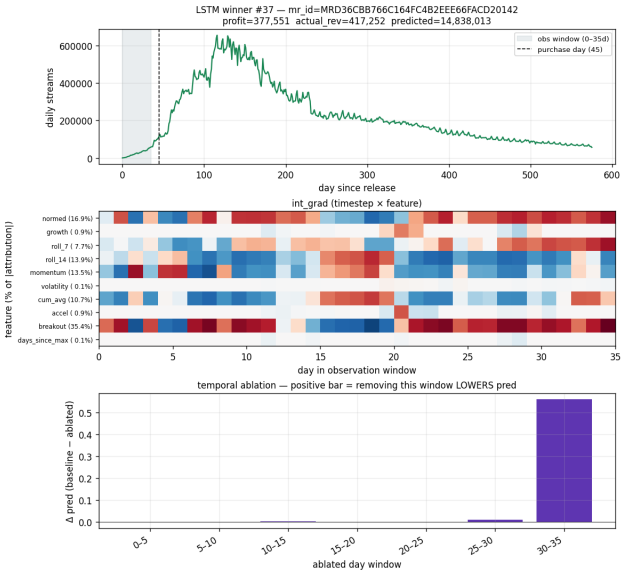

Figure 9: LSTM, profitable pick. Δprofit = +377,551; realised revenue 417,252.

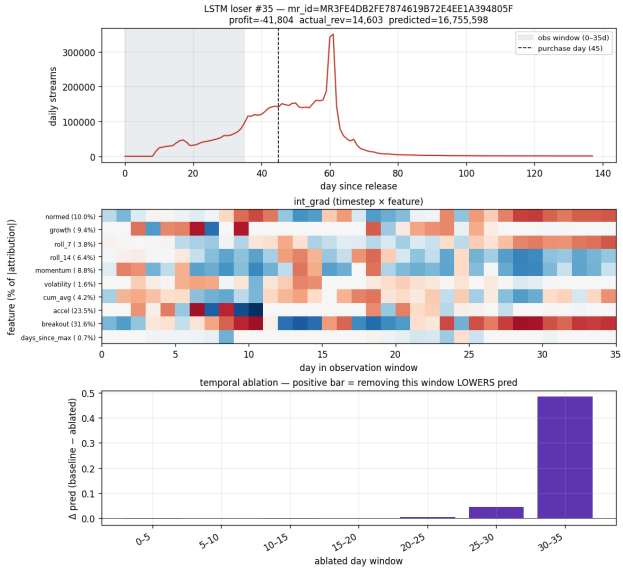

Figure 10: LSTM, loss-making pick. Δprofit = −41,804; realised revenue 14,603.