laurel

Customizing LLMs through Airflow

a ***cost sensitive*** approach to scalable model ***personalization***

13 AUG 24

laurel

Customizing LLMs through Airflow

and other ML models as well...

a ***cost sensitive*** approach to scalable model ***personalization***

13 AUG 24

00:

Moulay Zaidane Draidia

Data Scientist at Laurel

- Joined as founding member of AI team
- Model design, training, deployment, monitoring

laurel

00:

01: The Problem, the Company, and the Solution

02: Airflow at the core of ML training workflows

03: Airflow for Personalization

04: Airflow for Cost Sensitive Inference

laurel

01:

The **Problem**, the **Company**, and the **Solution**

laurel

Timekeeping is often cumbersome, highly manual, and time consuming

Professionals with high hourly fees are expected to be precise and accurate in their time tracking

Time is the most precious asset, understanding how it is allocated is vital for a business

- **Series B Company ($55M investment to date)**
- **Primary product is Automating Timekeeping for Time Professionals (Lawyers + Accountants).**
- **Shortlisted for ALM AI Product of the Year! ⚡**
- **Started with Global 100 law firms; In 2022 we partnered with Ernst & Young and officially entered the accounting vertical.**
- **Rebuilt our product from the ground up and launched 23 months**

THE SOLUTION

Bill against projects

Lawyers work on an average of 15 projects a day, sometimes exceeding 50

Track Hours worked

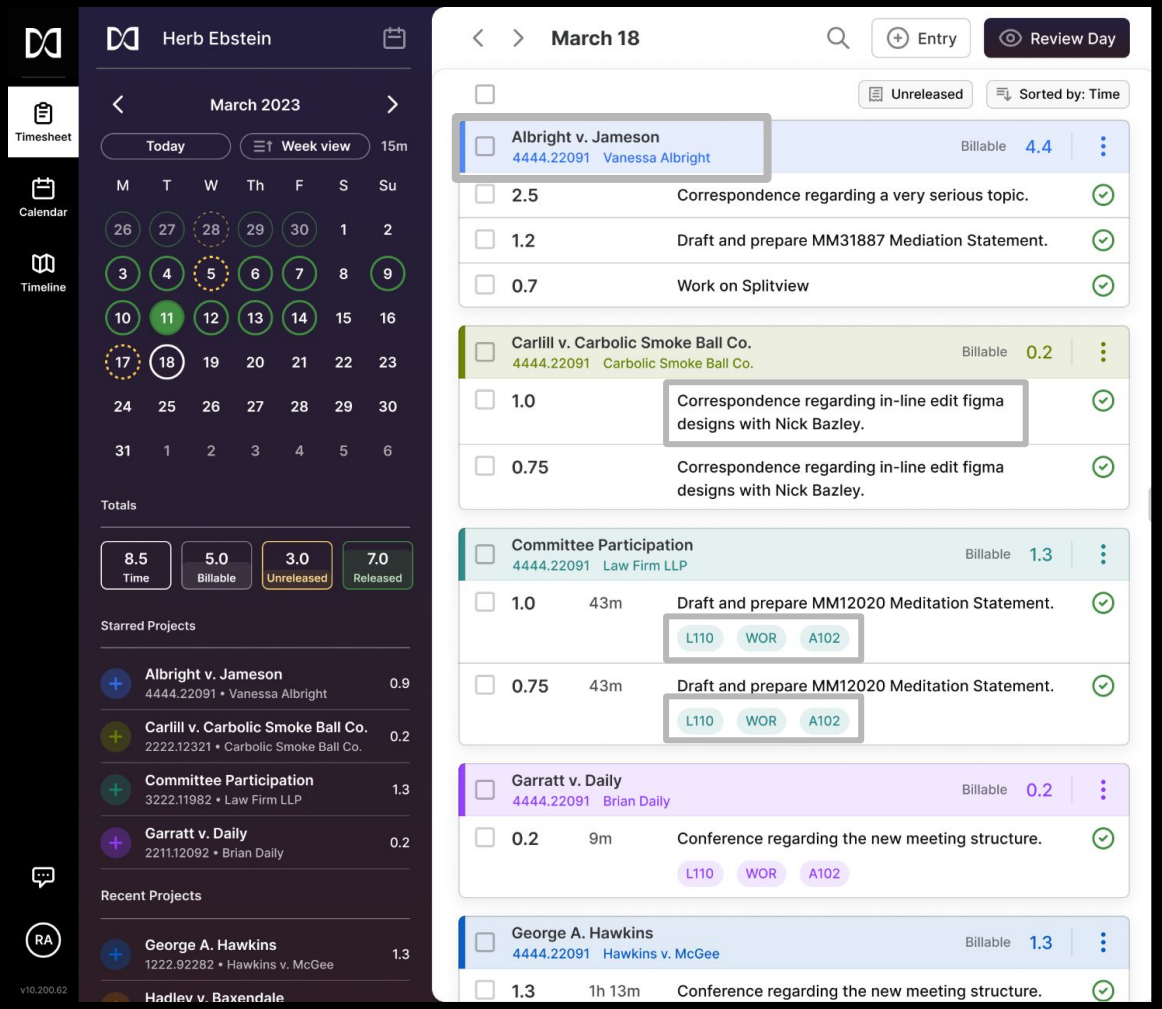
Lawyers are often expected to bill in 6 minute increments

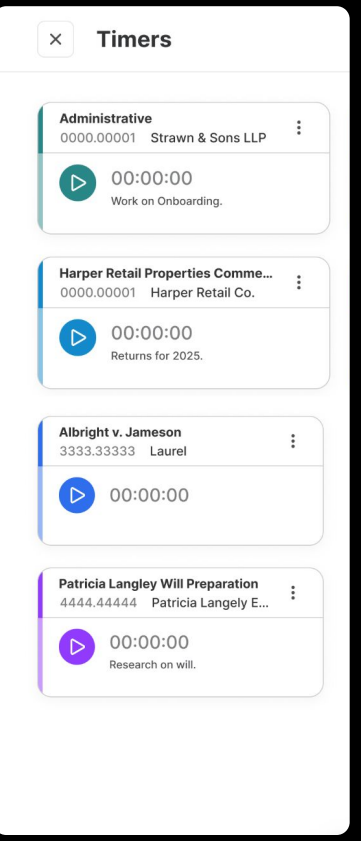
Describe the work

Company and clients issue specific guidelines dictating how to communicate the work performed

Label with codes

An extensive taxonomy of code is used to annotate each unit of work to facilitate downstream analysis and reporting





THE SOLUTION

Collect

Laurel Assistant accurately and exhaustively captures a knowledge worker's **entire** digital footprint

Group

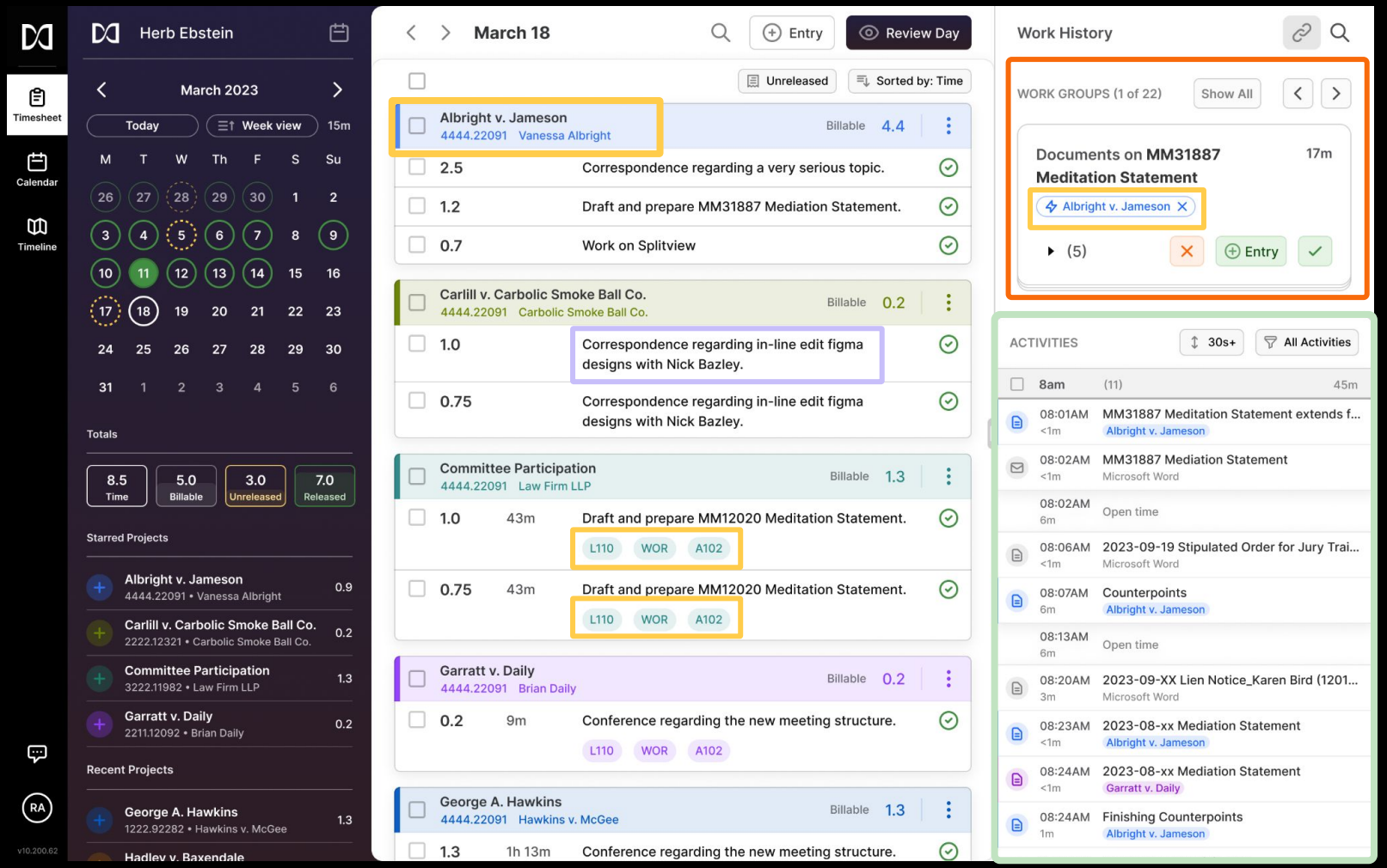
Laurel uses AI to intelligently group that digital activity into review ready time entries

Classify

Laurel predicts the client and project that the time entry to which the time entry should be billed

Summarize

Use GenAI to write ready-to-bill compliant narratives summarizing the work and minimizing rejection rates by clients.



02:

Airflow at the core of **ML training** workflows

laurel

Privacy

Legal data is particularly sensitive and requires strict control. We process confidential documents, contracts, correspondences etc...

Strict data isolation policies both across, and in some cases, within firms

Strong Data Drift

People work on a new projects for new clients every day. Laurel needs to adapt rapidly to each new initiative

Frequent model retraining is necessary

Varied user behavior

Every person has their unique work style

User level personalization goes a long way

In the critical path

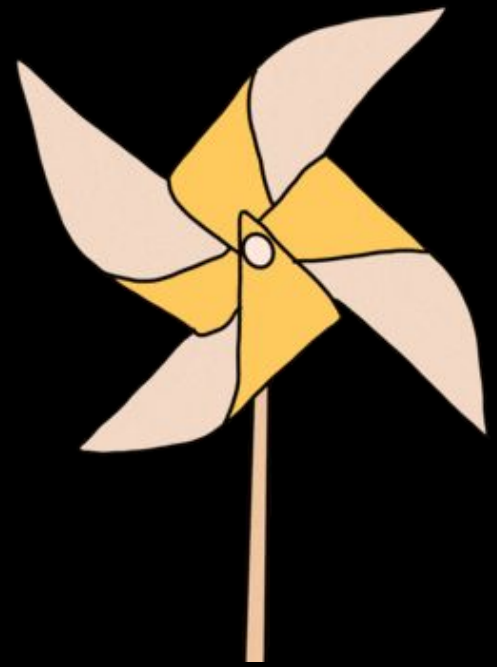
Users rely on Laurel to get paid. Trust in the accuracy of report is critical .

Accuracy, reliability, and responsiveness are key

Rebuilt our product in 5 months from the ground up, and launched 2 years ago

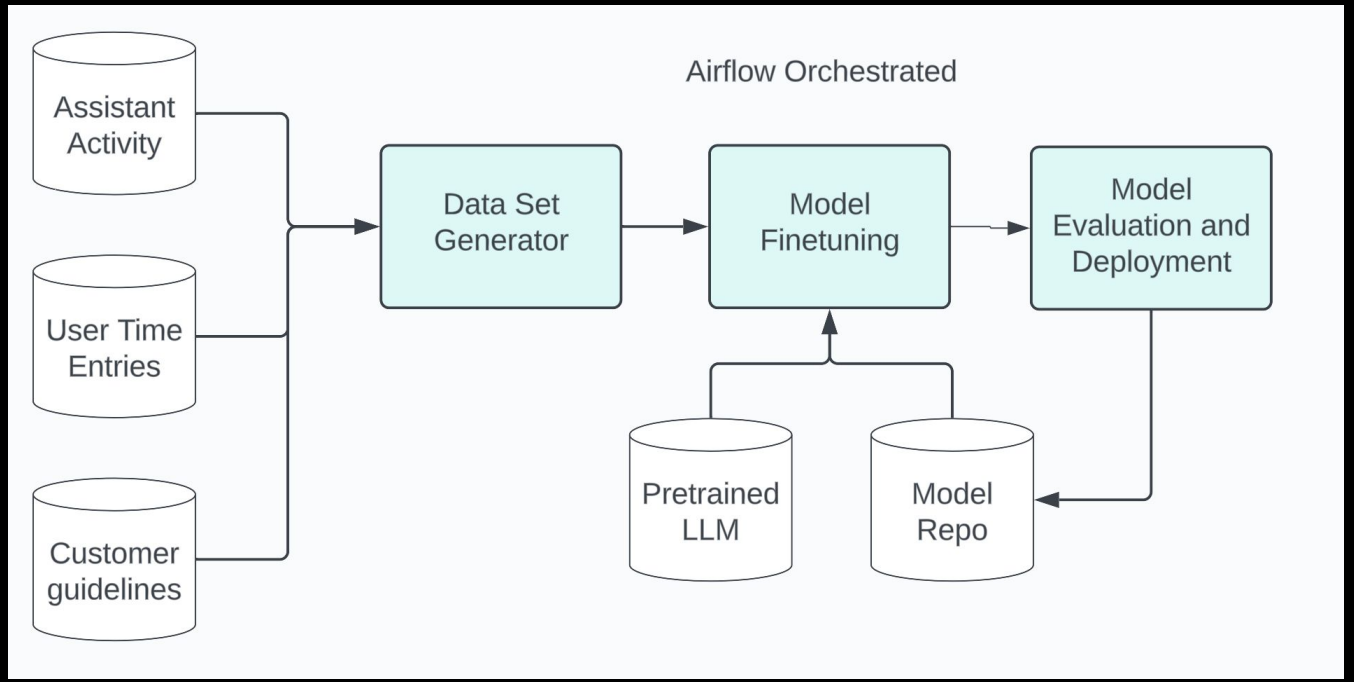
Airflow at the core of our ML use cases

- Airflow orchestration vs manual model training
- Manual deployments automated with Airflow
- Facilitated backfills feature engineering backfills
- Expensive simulation jobs run locally moved to airflow



Generalized model framework

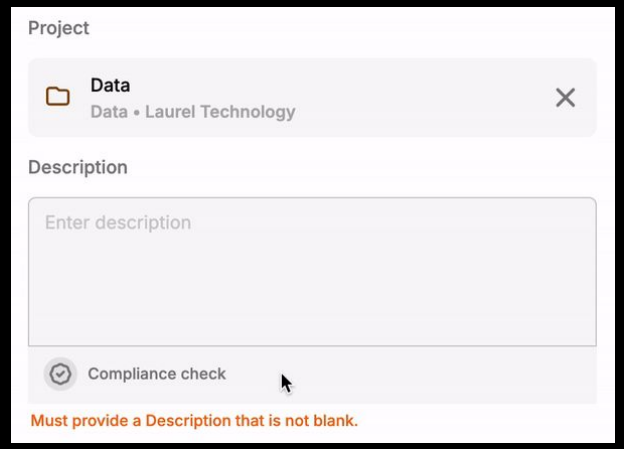
Keeping each step of our model generation process as modular as possible allows for rapid iteration and safe rollout of improvements



VERSATILE ACROSS MODELS

Autocompletion

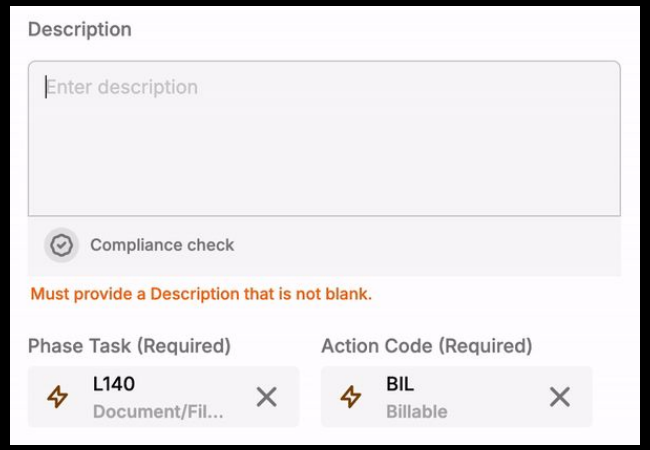
Bayesian Probability and Markov Chains



Daily DAG runs to store the conditional probabilities of a high likelihood phrases

Work Code prediction

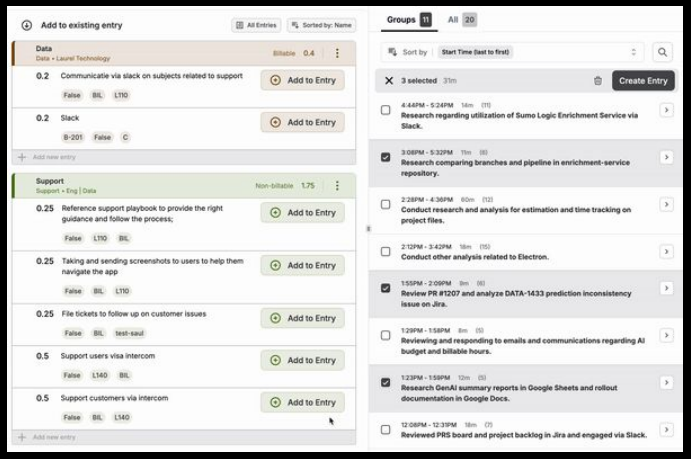
Siamese neural network



Weekly DAG runs to further train work code prediction, also triggered dynamically when new code taxonomy is ingested

Summarization

Fine tuned LLMs using RAG



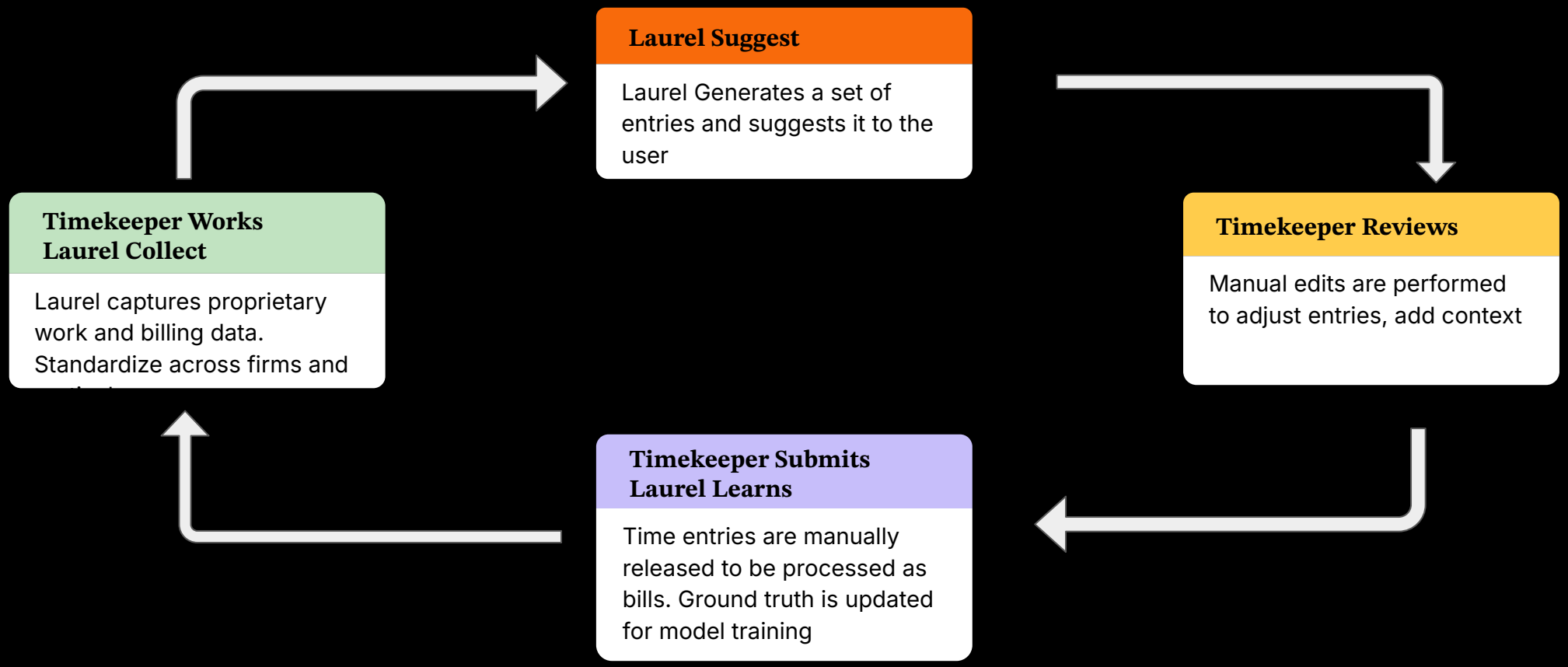
Orchestration of LLM fine tuning and iterative inference calibrated to usage patterns

03:

Airflow for **Personalization**

laurel

AI STRATEGY


Laurel Suggest
Laurel Generates a set of entries and suggests it to the user
Timekeeper Reviews
Manual edits are performed to adjust entries, add context
Timekeeper Submits Laurel Learns
Time entries are manually released to be processed as bills. Ground truth is updated for model training
Timekeeper Works Laurel Collect
Laurel captures proprietary work and billing data. Standardize across firms and

DAG parameters as Model Hyper Parameters

- How often should a dag be orchestrated?
- Which subset of data needs to be considered for each model run?
- Change retraining cadence or model architecture?

Efficient compute for Personalization

- Faster convergence when problem space is reduced
- While storage grows with more models, fine tuned artifacts can be compressed
- Efficient caching and dynamic model provisioning is critical to serve inference

Modeling Approach

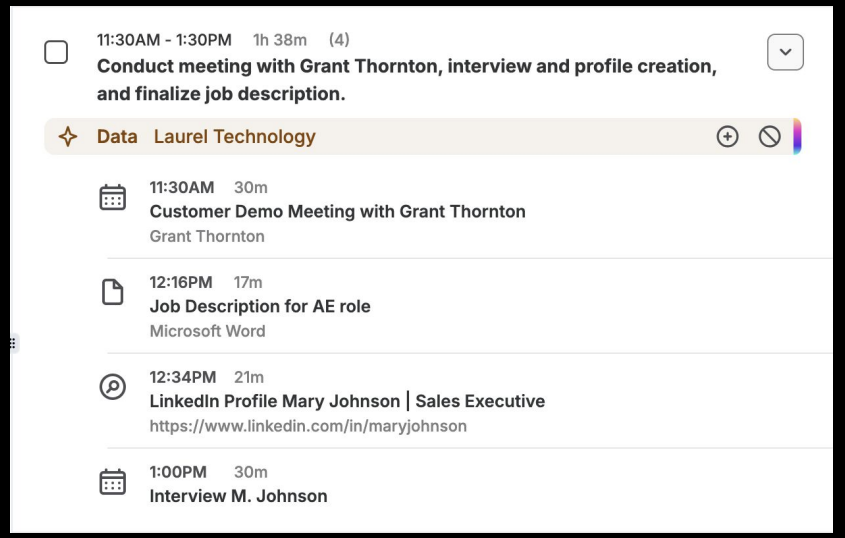
- **Similarity Classifier** determines whether two activities belong to the same entry
- Distance matrix as input to **Clustering algorithm**
- Prune outlier activities using Similarity Classifier
- **Confidence Classifier** to determine which groups to show case

Airflow's value

- **Multiple models for each user**
- Each model training can be orchestrated on the optimal cadence
- Facilitated experimentation
- Intuitive handling of model dependencies

Grouping

Semi-supervised clustering



Each user thinks and bills their work differently

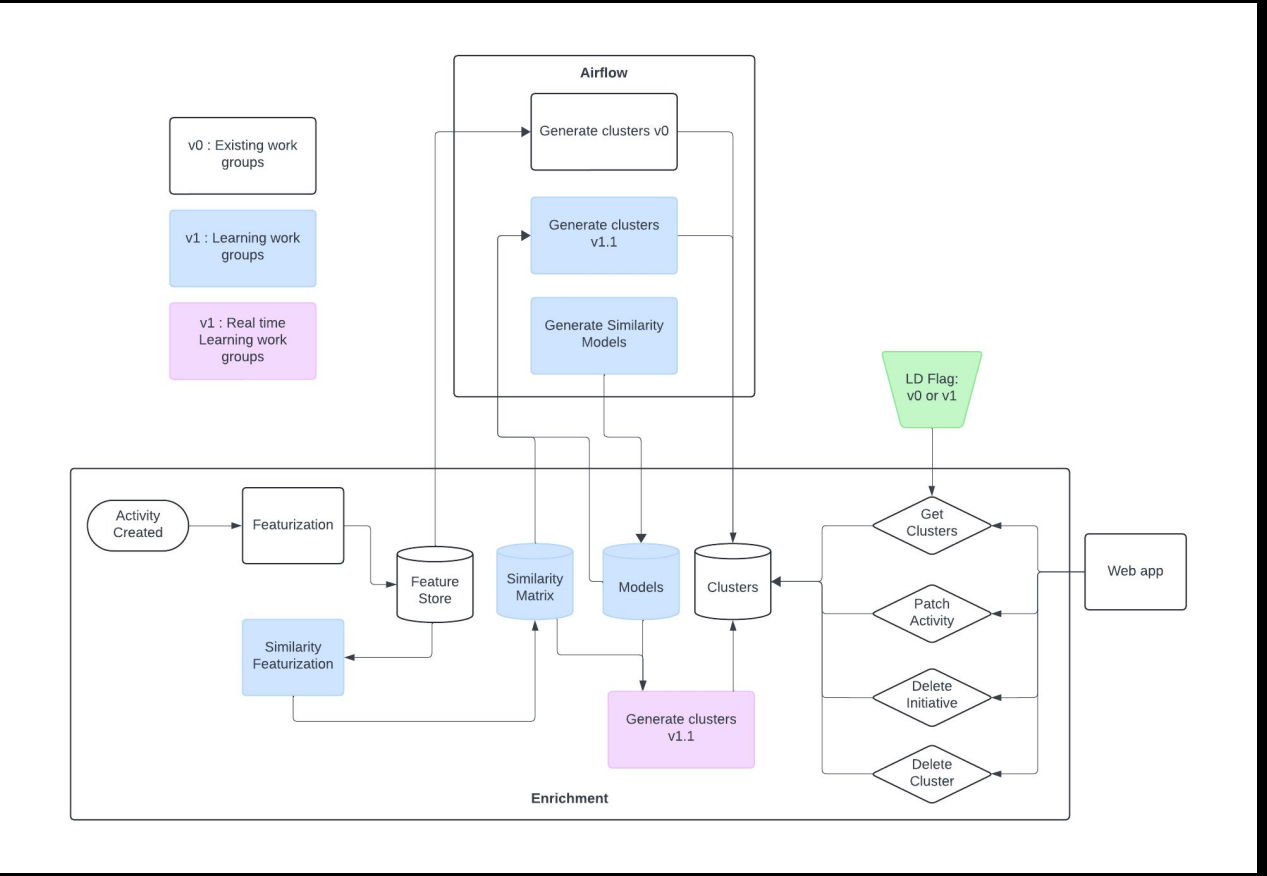
04:

Airflow for **Cost Sensitive Inference**

laurel

Work happens as a stream
Users bill on a cadence

- Keeping track of time contemporaneously can be a challenge
- Companies and clients set different expectations on release velocity (e.g. weekly, monthly...)
- Safely rollout by testing inference on DAGs before release synchronously



Work happens as a stream Users bill on a cadence

- Keeping track of time contemporaneously can be a challenge
- Companies and clients set different expectations on release velocity (e.g. weekly, monthly...)
- Safely rollout by testing inference on DAGs before release synchronously

Airflow orchestrates iterative inference

- Cheaper and faster models are used synchronously
- More expensive, slower and performant models are served on a cadence
- Offer the best experience without incurring unnecessary costs

Modeling Approach

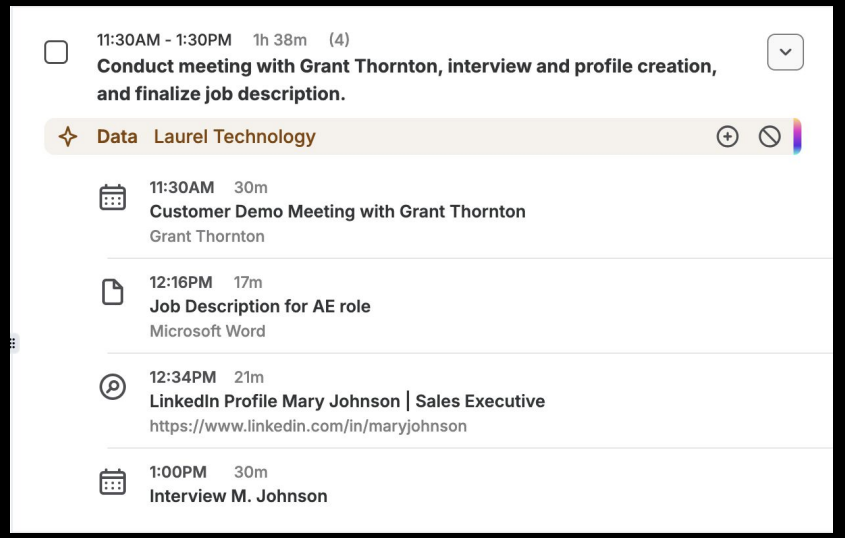
- Retrieval Augmented Generation (**RAG**) to generate prompt
  - firm and client guidelines
  - previously written summaries
  - new work activity

Airflow's value

- Significant cost savings
- Consistently provide descriptions
- Match inference schedule to billing schedule
- Facilitated experimentation - exploring fine tuning as a use case
- Provide a sense of improvement over time

Summary generation

LLM using RAG



Summarizing large amounts of documents, emails, work activity requires many tokens

05:

Summary

Airflow is a powerful ML orchestration engine

- Keeping each model orchestration step as a task / dag allows for **modularity**, **rapid iteration**, and **safe releases**
- DAG parameters as model **system hyperparameters**
- **Adapt inference cadence** to user behavior and needs
- Facilitated experimentation through backfills and simulations

laurel

Thank you for your time

We are hiring!