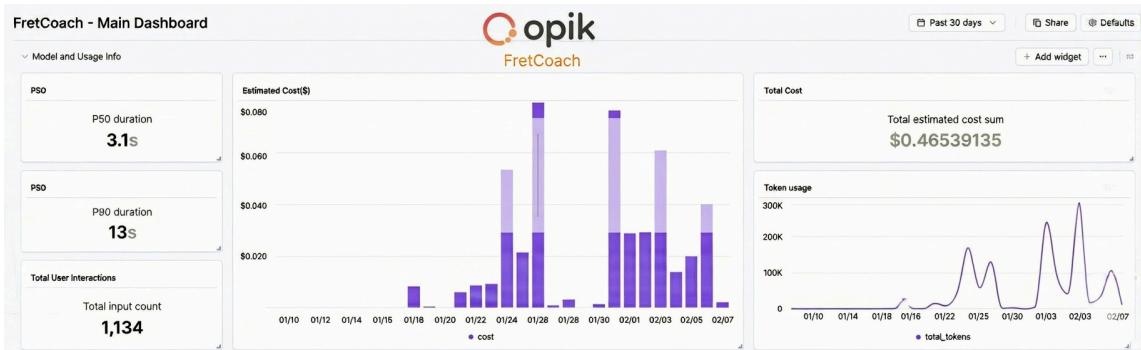


Opik Integration in FretCoach



Workspace: padmanabhan-r-7119 [Click To View Workspace](#)

Projects: FretCoach | FretCoach-Hub

Production Dashboard: [Click To View Dashboard](#)

Contents

- [Overview](#)
- [1. Traces with Metadata and Tags](#)
- [2. Thread Management](#)
- [3. Agent Graph Visualization](#)
- [4. Annotation Queues](#)
- [5. Datasets and Prompts](#)
- [6. Experiments and Custom Metrics](#)
- [7. Optimization Studio](#)
- [8. OpikAssist for Token Usage Optimization](#)
- [9. Project-Specific Configurations](#)
- [10. Online Evaluation](#)
- [11. Production Dashboard](#)
- [12. Alerts & Notifications](#)
- [13. Key Insights](#)

Overview

FretCoach traces three distinct AI coaching features — real-time live coaching, AI practice recommendations, and a web based chatbot agent (AI Coach). Opik provides the observability layer to monitor, evaluate, and continuously improve these features in production, giving us visibility into prompt quality, token costs, response latency, and AI coaching quality at scale.

Features Implemented

1. Traces with Metadata and Tags

All LLM calls are logged as traces in Opik with structured tags for filtering, organization and applying targeted online evaluation rules.

Hub Coach Chats:

- Tags: `ai-coach-chat`, `fretcoach-hub`, `from-hub-dashboard`, `gemini-2.5-flash`, `practice-plan`
- Tracks AI coach conversations in the web dashboard

The screenshot shows the Opik web interface. On the left is a sidebar with various project and monitoring options. The main area displays a trace analysis for 18 spans. A prominent feature is a `LangGraph` diagram at the top right, which illustrates the flow of the conversation with nodes for `start`, `agent`, `end`, and `tools`, along with a self-loop arrow on the `agent` node. Below the graph, there's a table of `Trace scores` with five rows, each with a key, score, and reason. The table includes columns for Key, Score, and Reason.

Key	Score	Reason
Hub Response Clarity	1	The response is clear, well-structured, ...
Hub Answer Correctness	1	The output provides a comprehensive ...
Hub Actionability	1	The response provides clear insights i...
Hub Data Groundedness	1	All factual claims in the output are fully...
Hub Context Usage Quality	1	The assistant effectively used the prov...

Hub coach chat traces with proper tags in Opik dashboard

AI Mode (Practice Recommendations):

- Tags: `fretcoach-core`, `gpt-4o-mini`, `ai-mode`, `fretcoach-studio`, `practice-recommendation`
- Tracks personalized practice recommendations

AI mode practice recommendation traces

Live AI Feedback in Session:

- Tags: `fretcoach-core`, `gpt-4o-mini`, `ai-mode`, `fretcoach-studio`, `live-feedback`
- Tracks real-time coaching feedback during practice
- TTS audio generation traced separately via `@track` decorator for independent failure tracking and latency monitoring

Live AI feedback traces during practice sessions

2. Thread Management

Structured `thread_id` to group related LLM calls and maintain conversation context.

Thread Naming Conventions:

- Hub Coach:** `hub-aicoach-chat-{timestamp}` - Groups all coach chat messages for a user
- AI Mode:** `{deployment}-ai-mode-{practice_id}` - Maintains thread across recommendations
- Live Feedback:** `{session_id}-live-aicoach-feedback` - Groups feedback within a practice session

A screenshot of the Opik interface showing a list of threads. Each thread entry includes a ID, first message, and last message. The threads listed are hub-aicoach-chat-1770250552190, hub-aicoach-chat-1770223524432, hub-aicoach-chat-1770217604585, hub-aicoach-chat-1770215363136, hub-aicoach-chat-1770215206179, e868ca6e-c204-45e8-83c0-0be930248505-live-aicoach-feedback, 6db7f5e0-1bdd-484d-9b1b-51059b3f8e90-live-aicoach-feedback, 0a7d4419-a110-4ff7-a8b6-ae56363283d7-live-aicoach-feedback, hub-aicoach-chat-1770035493514, hub-aicoach-chat-1770035435207, hub-aicoach-chat-1770035366500, and hub-aicoach-chat-1770035155059.

Thread IDs grouping related traces in Opik

3. Agent Graph Visualization

LangGraph execution flows visualized in Opik using `workflow.get_graph(xray=True)`.

Shows complete agent reasoning path: agent → tool calls (`execute_sql_query`, `get_database_schema`) → decision nodes → response.



LangGraph agent execution flow in Opik

4. Annotation Queues

Used annotation queues for human-in-the-loop evaluation of agent outputs.

Implementation:

- Custom feedback definitions for LLM output quality
- Manual review and rating using custom criteria
- Structured feedback collection for agent improvements

The screenshot shows a web-based annotation interface for AI Coach Chat Thread Quality Checks. At the top, there's a header bar with the URL 'comet.com/opik/padmanabhan-r-7119/annotation-queues/019c0544-cb75-7112-a0ef-5b13d2de588c?size=100&thread_height=medium&thread_filters=%5B%5D'. Below the header are several navigation and search buttons. The main area is titled 'AI Coach Chat Thread Quality Checks' and displays a table of annotation results. The table has columns for 'ID', 'First message', 'Last message', and 'Comments'. Each row contains a checkbox, the thread ID, the user's question, the AI's response, and a red circular icon with a number indicating the review status (e.g., 1, 2, 3). The interface includes various icons for filtering, sharing, and exporting.

ID	First message	Last message	Comments
thread-1769613376415	What should I practice today? I feel like doing some G# Minor. Can you make me a practice plan ?	{"messages": [{"content": "Hi! I'm your AI practice coach. I can help you analyze your performance, suggest exercises, and answer questions about your practice"}]	1 Very bad, no practice plan actually displayed.
thread-1769599262489	Show me my progress	{"messages": [{"content": "Hi! I'm your AI practice coach. I can help you analyze your performance, suggest exercises, and answer questions about your practice"}]	2 Add guard rails, as it is responding to random texts
thread-1769592723853	show me the scales that i have showed most improvement in	{"messages": [{"content": "Hi! I'm your AI practice coach. I can help you analyze your performance, suggest exercises, and answer questions about your practice"}]	3 Not many details
thread-176959257338	show me the scales that i have showed most improvement in	{"messages": [{"content": "Hi! I'm your AI practice coach. I can help you analyze your performance, suggest exercises, and answer questions about your practice"}]	-
thread-1769591135023	Show me my progress	{"messages": [{"content": "Hi! I'm your AI practice coach. I can help you analyze your performance, suggest exercises, and answer questions about your practice"}]	-

Annotation queue with reviewed LLM outputs

5. Datasets and Prompts

Created datasets and prompts for reproducible experiments and evaluations.

- Curated test cases from real user sessions
- Used in experiment runs for testing
- Version-controlled coaching prompt templates

Datasets created for experiment runs

Saved prompts

6. Experiments and Custom Metrics

Evaluated LLM performance using both default Opik metrics and custom-created metrics.

Default Metrics:

- Opik's built-in evaluation metrics for response quality

Custom Metrics:

- Domain-specific metrics tailored to guitar coaching context
- Measures coaching quality and relevance

comet.com/opik/padmanabhan-r-7119/experiments/019c2318-01d8-71a7-b2f7-c796b0577a6b/compare?experiments=%5B%019c2348-ef3f-7b71-97c5-88c73017e... ☆ 🔍

padmanabhan-r-7119 / Experiments / fretcoach-default-metrics-eval

fretcoach-default-metrics-eval

02/03/26 05:05 PM | fretCoach_live_ai_feedback_val_9 | Traces

Metrics: answer_relevance_metric (avg) 0.9044444444444444, context_precision_metric (avg) 0.7888888888888889, context_recall_metric (avg) 0.8388888888888889, hallucination_metric (avg) 0, levenshtein_r

Experiment items Configuration Feedback scores

Experiment items are individual evaluations that connect a dataset sample with its LLM output, feedback scores, and trace. Read more

Search dataset items Compare

ID (Dataset item)	Dataset	input	Evaluation task	input	output	reference	Duration	Total tokens
019c2325-43c2-73b...	Your pitch accuracy is strong, but timing stability is lacking. Slow	Enabled metrics: Pitch Accuracy, Timing Stability Pitch 89%, Timing	0: "You are a direct guitar coach giving	Enabled metrics: Great pitch accu...	Your pitch accur...	68.4s	p50 67.6s avg 13042.667	12,70
019c2325-43c2-73b...	Timing is solid, but scale conformity is weak. Focus on exploring different	Enabled metrics: Pitch Accuracy, Scale Conformity, Timing Stability	0: "You are a direct guitar coach giving	Enabled metrics: Great timing stab...	Timing is solid, b...	70.3s	14,86	
019c2325-43c2-73b...	Excellent pitch accuracy, but timing stability could improve.	Enabled metrics: Pitch Accuracy, Timing Stability Pitch 97%, Timing	0: "You are a direct guitar coach giving	Enabled metrics: Great job on pitc...	Excellent pitch a...	71.5s	13,56	

Experiments with default Opik metrics

comet.com/opik/padmanabhan-r-7119/experiments/019c2318-01d8-71a7-b2f7-c796b0577a6b/compare?experiments=%5B%019c2345-2271-71cd-a540-ab7b48b9... ☆ 🔍

padmanabhan-r-7119 / Experiments / fretcoach-coaching-feedback-eval

fretcoach-coaching-feedback-eval

02/03/26 05:01 PM | fretCoach_live_ai_feedback_val_9 | Traces

Metrics: coaching_quality (avg) 0.7888888888888889

Experiment items Configuration Feedback scores

Experiment items are individual evaluations that connect a dataset sample with its LLM output, feedback scores, and trace. Read more

Search dataset items Compare

ID (Dataset item)	Dataset	input	Evaluation task	input	output	reference	Duration	Total tokens	Estimated cost
019c2325-43c2-73b...	Your pitch accuracy is strong, but timing stability is lacking. Slow	Enabled metrics: Pitch Accuracy, Timing Stability Pitch 89%, Timing	Enabled metrics: Great pitch accu...	Your pitch accur...	4.1s	p50 3.1s avg 0	0		
019c2325-43c2-73b...	Timing is solid, but scale conformity is weak. Focus on exploring different	Enabled metrics: Pitch Accuracy, Scale Conformity, Timing Stability	Enabled metrics: Great timing stab...	Timing is solid, b...	3.1s	-	-		
019c2325-43c2-73b...	Excellent pitch accuracy, but timing stability could improve.	Enabled metrics: Pitch Accuracy, Timing Stability Pitch 97%, Timing	Enabled metrics: Great job on pitc...	Excellent pitch a...	3.2s	-	-		
019c2325-43c2-73b...	Timing is decent,	Enabled metrics:	Enabled metrics: Good timing, but ...	Timing is decent,...	3.5s	-	-		

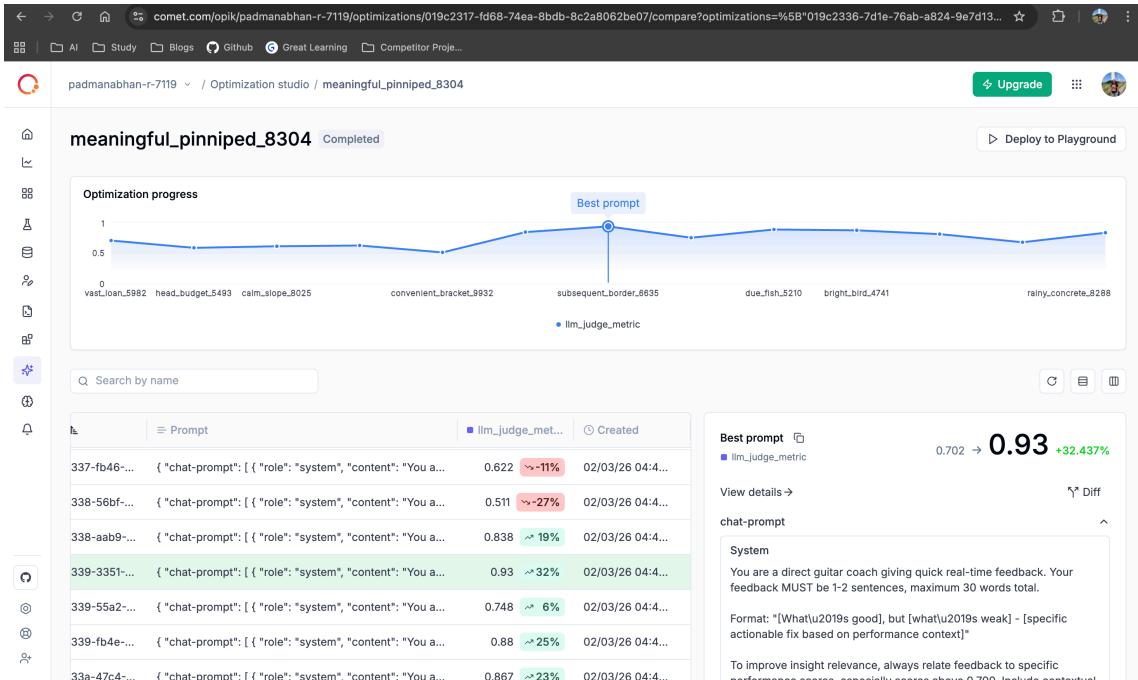
Experiments with custom metric

7. Optimization Studio

Used Optimization Studio with HRPO (Hierarchical Reflective Prompt Optimizer) to improve the prompt used in the live feedback module.

Results:

- 32% increase in `llm_judge_metric` custom metric
- Improved coaching feedback quality
- Optimized for better real-time guidance



Optimization Studio results for live feedback prompt

8. OpikAssist for Token Usage Optimization

Used OpikAssist to analyze traces and optimize token usage for hub coach chats.

Problem Identified:

- Excessive token usage (4,877 tokens) and long duration (7,873 ms)
- Lengthy prompts with redundant context and full SQL data

Actions Taken:

- Refined system and user prompts based on OpikAssist suggestions
- Streamlined SQL result formatting to essential data only
- Removed redundant context and consolidated guidelines

Results:

- Significant token usage reduction
- Improved response latency
- Better cost-performance ratio

The screenshot shows the Opik platform's interface. On the left, a sidebar navigation includes Home, Dashboards, Observability, Projects (selected), Evaluation, Experiments, Datasets, Annotation queue, Prompt engineering, Prompt library, Playground, Optimization (selected), and Production. A message bar at the top right says "OpikAssist Beta". The main area displays a "Trace - 8 spans" section with a tree view of spans categorized by agent, tools, and another agent. An "Output" panel shows a message from the AI: "Alright Paddy, let's take a look at your recent progress! Based on your last 20 practice sessions, I've noticed some interesting trends: Pitch Accuracy: Your pitch accuracy has generally been quite strong, often hitting above 70%, with some excellent sessions in A Minor reaching up to 86%! Keep up the great work on hitting those notes precisely. Timing Stability: You're showing good consistency with your timing, often staying in the 70-80% range, and even hitting an impressive 95% in one of your D Minor sessions. This indicates a solid sense of rhythm. Scale Conformity: This is an area where we see a bit more fluctuation. While you have some great sessions, especially in A Minor (up to 62%), there are also sessions where scale conformity is lower, particularly when exploring new scales like G Major or D Minor. This suggests that understanding and consistently applying the notes within a chosen scale could be a good focus area for you. You've been exploring a good variety of scales including E Major, A Minor, G Major, D Minor, C# Minor, C Major, and E Minor. It's great to see you challenging yourself with different keys! A performance trend chart will appear below automatically, giving you a visual overview of these metrics over time. What do you think of these insights, Paddy? Would you like a practice plan to help improve your scale conformity, or perhaps focus on something else?" Below this, a message input field says "Type your message" and a note says "Our AI assistant allows you to analyze trace and spans data (which may include personal or sensitive information) using a generative AI model via OpenAI, L.L.C.".

OpikAssist analyzing trace for token usage optimization

9. Project-Specific Configurations

Configured custom feedback definitions and AI providers for comprehensive evaluation.

Feedback Definitions:

- Custom fields for manual LLM output rating
- Human-in-the-loop feedback on traces
- Categorical ratings: "AI Coach Conversation Rating" and "User Feedback"

AI Providers:

- Perplexity's Sonar Pro for automated evaluations
- OpenRouter models for diverse evaluation perspectives
- Enhanced evaluation capabilities beyond default Opik models

The screenshot shows the "Feedback definitions" page. At the top, tabs include "Feedback definitions" (selected), AI Providers, Workspace preferences, and Members. A note says "Create custom fields to manually rate LLM outputs. Use them to collect structured feedback and track quality over time." Below is a search bar and a "Create new feedback definition" button. A table lists three feedback definitions: "Feedback score" (Type: Categorical, Values: Average, Bad, Excellent, Good, Very Good), "AI Coach Conversation Rating" (Type: Categorical, Values: Average, Bad, Excellent, Good, Very Good), and "User Feedback" (Type: Categorical, Values: 🌟, 🌟).

Custom feedback definitions for manual trace ratings

Feedback definitions	AI Providers	Workspace preferences	Members
ⓘ Connect AI providers to test prompts, preview model responses, and score traces using online evaluation rules in the Playground. ×			
<input type="text" value="Q Search by name"/>			Add configuration
≡ Name	≡ URL	≡ Created	≡ Provider
Perplexity	https://api.perplexity.ai	01/28/26 08:36 PM	▼ Perplexity ...
OPENROUTER_API_KEY	-	01/28/26 08:33 PM	◀ OpenRouter ...
OPIK_FREE_MODEL_API_KEY	-		○ Opik Read-only provider

Custom AI providers configured for automated evaluations

10. Online Evaluation

Configured **11 online evaluation rules** to automatically score production traces using LLM-as-a-Judge metrics.

Purpose:

- Real-time quality monitoring of AI responses in production
- Automatic evaluation without manual review
- Early detection of performance degradation or quality issues

Rules Overview:

Hub Coach (7 rules):

1. hub_answer_correctness - Validates factual accuracy
2. hub_data_groundedness - Ensures grounding in database context
3. hub_context_usage_quality - Checks effective use of retrieved data
4. hub_actionability - Measures actionable guidance
5. hub_response_clarity - Evaluates readability
6. hub_conversational_coherence - Tracks conversation flow (thread-level)
7. hub_user_frustration_score - Detects user frustration (thread-level)

Studio AI Mode (4 rules):

8. studio_practice_recommendation_alignment - Validates goal alignment
9. studio_immediate_actionability - Ensures executable recommendations
10. studio_live_coach_feedback_quality - Measures real-time coaching quality
11. studio_live_feedback_effectiveness - Tracks session improvement (thread-level)

[View complete rule prompts and variable mappings →](#)

The screenshot shows a web-based interface for managing online evaluation rules. At the top, there's a header bar with a logo, a search bar, and a 'Upgrade' button. Below the header is a sidebar with various icons. The main area is titled 'Online evaluation' and contains a sub-header 'Automatically score your production traces by defining LLM-as-a-Judge or code metrics. [Read more](#)'. A search bar labeled 'Search by ID' is followed by a table with 11 rows of data. The columns in the table are: Name, Projects, Scope, Status, and Show logs. Each row has a checkbox in the first column and a '...' button in the last column. The table footer indicates 'Showing 1-10 of 11'.

	Name	Projects	Scope	Status	Show logs
<input type="checkbox"/>	hub_answer_correctness	FretCoach, FretCoach-Hub	Trace	• Enabled	Show logs
<input type="checkbox"/>	hub_data_groundedness	FretCoach, FretCoach-Hub	Trace	• Enabled	Show logs
<input type="checkbox"/>	hub_context_usage_quality	FretCoach, FretCoach-Hub	Trace	• Enabled	Show logs
<input type="checkbox"/>	hub_actionability	FretCoach, FretCoach-Hub	Trace	• Enabled	Show logs
<input type="checkbox"/>	hub_response_clarity	FretCoach, FretCoach-Hub	Trace	• Enabled	Show logs
<input type="checkbox"/>	studio_practice_recommendation_alignment	FretCoach	Trace	• Enabled	Show logs
<input type="checkbox"/>	studio_immediate_actionability	FretCoach	Trace	• Enabled	Show logs
<input type="checkbox"/>	studio_live_coach_feedback_quality	FretCoach	Trace	• Enabled	Show logs
<input type="checkbox"/>	hub_conversational_coherence	FretCoach, FretCoach-Hub	Thread	• Enabled	Show logs
<input type="checkbox"/>	studio_live_feedback_effectiveness	FretCoach, FretCoach-Hub	Thread	• Enabled	Show logs

Online evaluation rules dashboard (1-10 of 11)

This screenshot shows the same 'Online evaluation rules dashboard' as the previous one, but it displays only the last 11 rules from the total of 22. The table structure is identical, with columns for Name, Projects, Scope, Status, and Show logs. The table footer indicates 'Showing 11-21 of 22'.

	Name	Projects	Scope	Status	Show logs
<input type="checkbox"/>	hub_user_frustration_score	FretCoach, FretCoach-Hub	Thread	• Enabled	Show logs

Online evaluation rules dashboard (11 of 11)

11. Production Dashboard

A real-time dashboard monitoring key AI quality metrics across FretCoach's Studio and Hub applications.

[View Live Dashboard](#)

Dashboard Structure:

The dashboard displays 7 core metrics organized by application:

FretCoach - Studio and Portable (Core Functionality)

Evaluates practice recommendations and real-time coaching effectiveness.

Metric	Range	What the Score Means
Live Coach Feedback Quality	1 – 4	1 = Bad, 2 = Good, 3 = Very Good, 4 = Excellent
Practice Recommendation Alignment	0.0 – 1.0	1.0 = Aligned with the player's weaknesses; 0.0 = No alignment

⚡ Practice Recommendation - Immediate Actionability	0.0 – 1.0	1.0 = Recommendation fully actionable; 0.0 = Not actionable
--	------------------	---

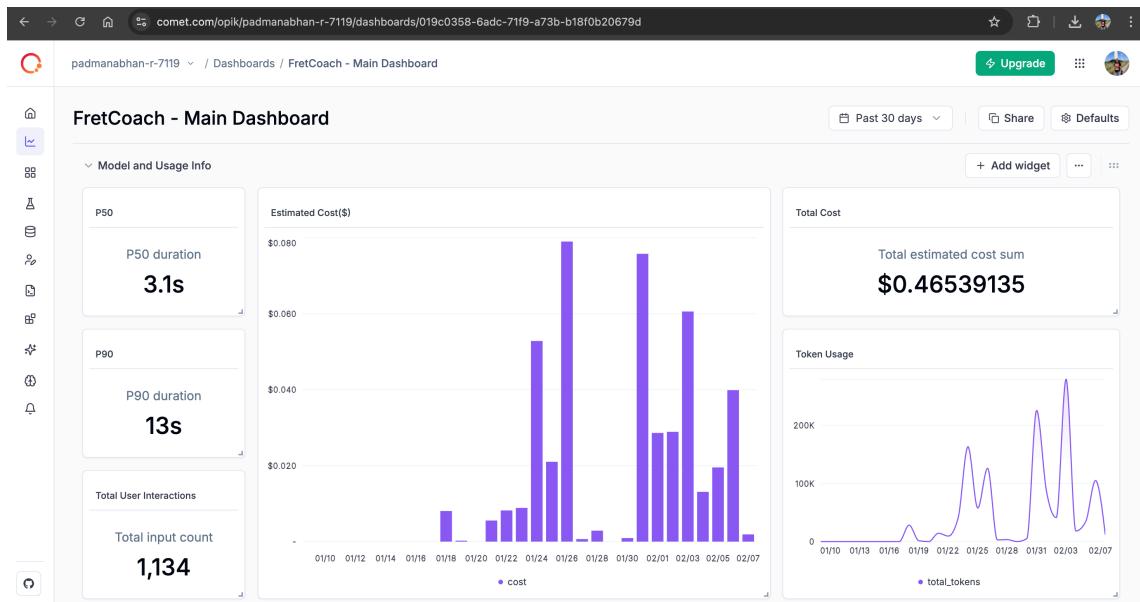
🧠 FretCoach Hub (Web)

Measures how well the Hub understands, answers, and guides users.

Metric	Range	What the Score Means
🌟 Response Clarity	0.0 – 1.0	1.0 = Clear and structured response; 0.0 = Poor response
🔗 Context Usage Quality	0.0 – 1.0	1.0 = Context effectively used; 0.0 = Poor usage of context
🎯 Actionability	0.0 – 1.0	1.0 = Clear, executable next steps; 0.0 = Vague or non-actionable
📊 Data Groundedness	0.0 – 1.0	1.0 = Supported by user's practice data; 0.0 = Weak grounding

Features:

- Real-time metric averages calculated from production traces
- Automatic updates as new LLM calls are evaluated
- Clear visibility into AI quality across different use cases
- Easy identification of performance degradation



Dashboard overview - Model and usage statistics

The screenshot shows a detailed view of the FretCoach - Main Dashboard. At the top, there's a navigation bar with a user icon, a dropdown menu, and links for 'Dashboards' and 'FretCoach - Main Dashboard'. On the right side of the header are buttons for 'Upgrade', 'Share', and 'Defaults'. Below the header, there's a search bar with a dropdown for 'Past 30 days', a 'Share' button, and a 'Defaults' button. A 'FretCoach - AI Evaluation Metrics Info' section is expanded, titled 'FretCoach - AI Evaluation Metrics'. It contains two tables: one for 'FretCoach - Studio and Portable (Core Functionality)' and another for 'FretCoach Hub (Web)'. Both tables provide metric details, ranges, and score meanings.

AI Evaluation Metrics documentation embedded in dashboard

This screenshot shows the live production dashboard for FretCoach. The layout is similar to the main dashboard, with a sidebar on the left and various sections on the right. The 'FretCoach - Main Dashboard' section is visible at the top. Below it, the 'FretCoach - Studio and Portable (Core Functionality)' section is expanded, showing three cards with average scores: 'Average Live Coach Feedback Quality' (3.432), 'Average Practice Recommendation Alignment' (0.884), and 'Average Practice Recommendation - Immediate Actionability' (0.947). The 'FretCoach Hub (Web)' section is also expanded, showing four cards with average scores: 'Average Hub Response Clarity' (0.972), 'Average Hub Context Usage Quality' (0.97), 'Average Hub Actionability' (0.994), and 'Average Hub Data Groundedness' (0.834). A 'Add section' button is located at the bottom center of the dashboard area.

Live production dashboard with real-time AI quality metrics

12. Alerts & Notifications

Configured Slack alerts to proactively monitor AI quality and system health in production.

Setup:

- Created a dedicated Slack channel: #opik-alerts
- Integrated Opik with Slack using webhook configuration

- Configured alerts for critical metrics and system errors

Alert Types:

1. Trace Errors Threshold

- Trigger:** When trace error count exceeds 10 in the last 30 minutes
- Purpose:** Detect system failures or integration issues

2. Feedback Score Thresholds

- Trigger:** When average metric scores fall below 0.6 in the last 30 minutes
- Monitored Metrics:**
 - Hub Response Clarity < 0.6
 - Hub Data Groundedness < 0.6
 - Hub Context Usage Quality < 0.6
 - Hub Answer Correctness < 0.6
 - Hub Actionability < 0.6
- Purpose:** Early detection of AI quality degradation

3. Latency Alerts

- Trigger:** When average latency exceeds 3 seconds in the last 30 minutes
- Purpose:** Monitor response time performance and identify slowdowns

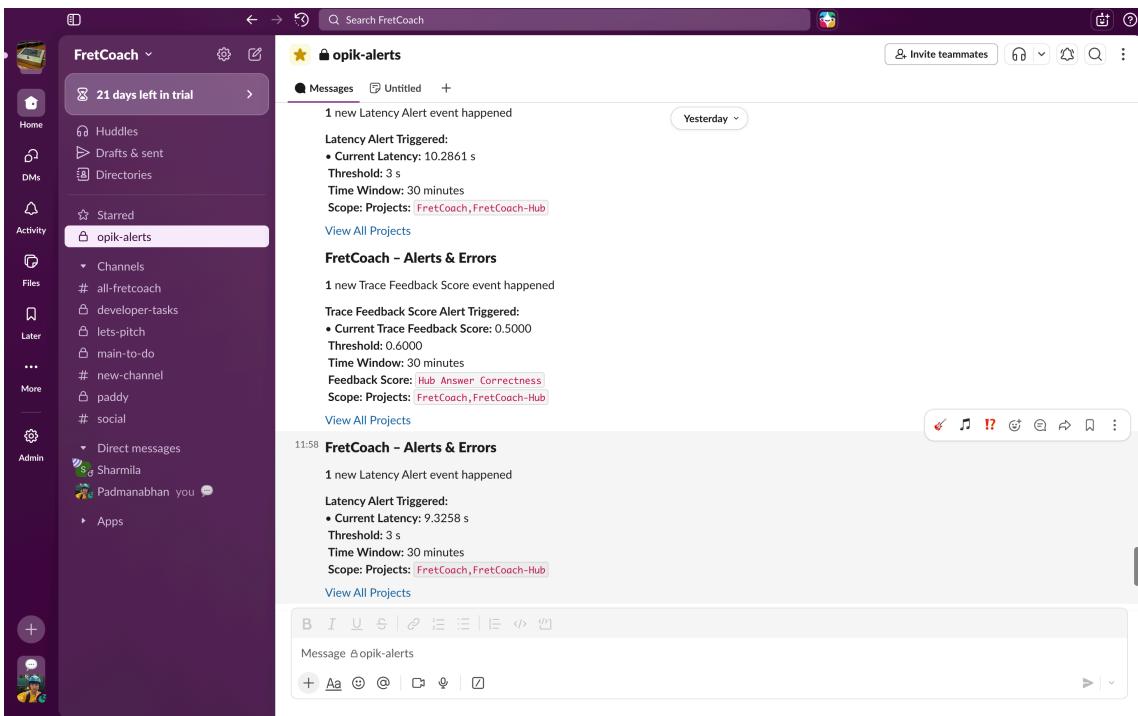
Benefits:

- Proactive issue detection before users report problems
- Real-time visibility into production AI quality
- Team-wide awareness through Slack notifications
- Quick response to quality degradation or system errors

```

    "blocks": [
      {
        "type": "header",
        "text": {
          "type": "plain_text",
          "text": "FretCoach - Alerts & Errors"
        }
      },
      {
        "type": "section",
        "text": {
          "type": "mrkdwn",
          "text": "*Trace Errors Alert Triggered*\n\n*Current Trace Errors*: 15\n*Threshold*: 10\n*Time Window*: 1 hour\n*Scope*: *Projects*: *Demo Project, Default Project*\n\nhttp://localhost:5173/demo_workspace_name/projects/view_all_projects"
        }
      }
    ]
  }
}
  
```

Alert configuration in Opik dashboard



Real-time alerts delivered to Slack #opik-alerts channel

13. Key Insights Gained from Production Observability

Opik traces surfaced actionable improvements that directly improved FretCoach's quality and performance:

Insight	Discovery	Action Taken	Result
Prompt verbosity	Live coach prompts were verbose, causing slow responses	Tightened prompt to "1-2 sentences, max 30 words" constraint	Significantly faster responses, more focused feedback
TTS latency spike	TTS taking longer than expected on some calls	Implemented singleton audio player instance to prevent concurrent playback	Consistent TTS latency, no audio overlap
Prompt optimization	Live feedback prompt quality measured via <code>llm_judge_metric</code>	Used Optimization Studio (HRPO) to refine the prompt	32% increase in live coaching response quality
Fallback model visibility	~15% of Hub Chat requests hit Gemini rate limits	Confirmed MiniMax fallback working seamlessly, kept hybrid approach	Zero user-facing errors on rate limit
Token cost patterns	Different features had very different token footprints	Targeted gpt-40-mini for cost-sensitive real-time features	Optimized cost-performance ratio per feature

Ongoing: All 11 online evaluation rules continue to run in production, monitoring AI quality across both Studio and Hub with automatic alerts when scores degrade.