

Opik Integration in FretCoach

Workspace: padmanabhan-r-7119 Projects: FretCoach | FretCoach-Hub Link: [View Workspace →](#)

Overview

FretCoach traces three distinct AI coaching features — real-time live coaching, AI practice recommendations, and a natural language web dashboard agent. 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 and organization.

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 dashboard interface. On the left, there's a sidebar with various project and tool options like Home, Dashboards, Observability, Projects, Experiments, Datasets, Annotation queue, Prompt engineering, Prompt library, Playground, Optimization, Optimization stud, Production, Online evaluation, Alerts, Star (17.5K), Configuration, Support hub, and Invite a teammate. The 'Projects' option is currently selected. The main content area has a header 'Trace - 18 spans' and a tree view of the trace structure. It includes sections for 'LangGraph', 'agent', 'should_continue', 'tools', and 'execute_sql_query'. To the right, there's a detailed view of the trace flow, showing nodes for 'start', 'agent', 'end', and 'tools' with arrows indicating the flow. Below this, there are tabs for 'Input/Output', 'Feedback scores (5)', and 'Metadata'. The 'Feedback scores' tab is active, showing a table with five rows of data. The 'Metadata' tab shows a table of trace scores with columns for 'Key', 'Score', and 'Reason'.

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

Trace - 2 spans

RunnableSequence

3.1s # 2084 1997/87 <\$0.001 2 5

ChatOpenAI

3.1s # 2084 1997/87 <\$0.001 openai gpt-4o-mini-2024-07-18

RunnableLambda

0.00ts

RunnableSequence

3.1s # 2084 1997/87 <\$0.001 2 2

ai-mode fretcoach-core fretcoach-studio gpt-4o-mini practice-recommendation

Details Feedback scores

Input

Pretty

You are an AI guitar coach. Based on the practice history below, recommend a practice session.

PRACTICE HISTORY:

- Total sessions: 8
- Average pitch accuracy: 71.6%
- Average scale conformity: 21.1%
- Average timing stability: 56.5%
- Weakest area: scale

RECENTLY PRACTICED SCALES:

```
[  
{  
  "scale_name": "A Minor",  
  "scale_type": "pentatonic",  
  "times_practiced": 8,  
  "avg_pitch": 0.71617078055838,  
  "avg_scale": 0.210802486501322,  
  "avg_timing": 0.565457327091952,  
  "last_practiced": "2026-01-30T23:11:58.958641"  
}  
]
```

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

The screenshot shows the Comet Debugger interface with the following details:

- Trace - 1 spans**:
 - ChatOpenAI**:
 - 2s # 389 361/28 <\$0.001 1 5
 - ChatOpenAI**:
 - 2s # 389 361/28 <\$0.001 5 @ openai gpt-4o-mini-2024-07-18
- Details**:
 - Input**:

Pretty

Enabled metrics: Pitch Accuracy, Timing Stability
Pitch 95%, Timing 21%
Strongest: Pitch Accuracy (95%)
Weakest: Timing Stability (21%)

Give 1-2 sentences (max 30 words) - what's good, what's weak, specific actionable fix:
 - Output**:

Pretty

Your pitch accuracy is outstanding at 95%, but timing stability is lacking at 21%. Slow down and count to improve your rhythm consistency.
 - Metadata**:

YAML

```
1 v providers:
```

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-{user_id}` - 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 unique ID, a timestamped message, and a JSON snippet of the message content. The threads are categorized by their purpose, such as AI practice or live feedback.

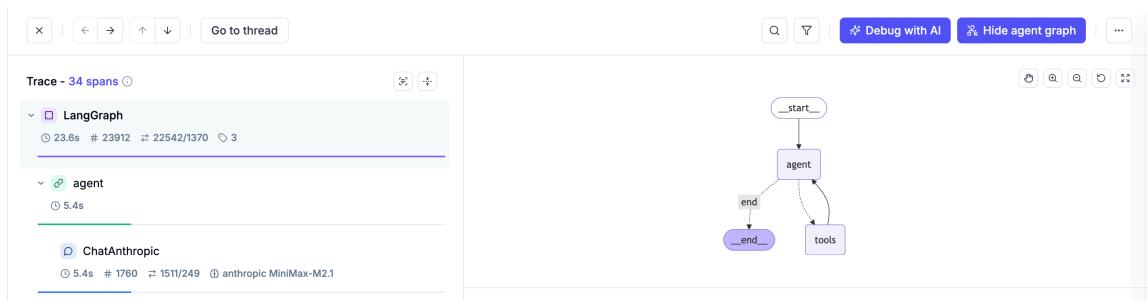
ID	Message	Content Snippet
hub-aicoach-chat-1770250552190	How am I doing compared to average?	{"messages": [{"content": "Hi! I'm your AI practice co..."}]
hub-aicoach-chat-1770223524432	How am I doing compared to average? And what day is ...	{"messages": [{"content": "Hi! I'm your AI practice co..."}]
hub-aicoach-chat-1770217604585	How am I doing compared to average?	{"messages": [{"content": "Hi! I'm your AI practice co..."}]
hub-aicoach-chat-1770215363136	What should I practice today?	```json { "focus_area": "Scale Conformity", "current...`
hub-aicoach-chat-1770215206179	What should I practice today?	{"messages": [{"content": "Hi! I'm your AI practice co..."}]
e868ca6e-c204-45e8-83c0-0be930248505-live-aicoach-feedback	Enabled metrics: Pitch Accuracy, Scale Conformity, T...	Timing is spot on at 100%, but scale conformity is...
6db7f5e0-1bdd-484d-9b1b-51059b3f8e90-live-aicoach-feedback	Enabled metrics: Pitch Accuracy, Scale Conformity, T...	Timing is excellent at 97%, but scale conformity is...
0a7d4419-a110-4ff7-a8b6-ae56363283d7-live-aicoach-feedback	Enabled metrics: Pitch Accuracy, Scale Conformity, T...	Timing is excellent at 98%, but scale conformity is...
hub-aicoach-chat-1770035493514	What should I practice today?	{"messages": [{"content": "Hi! I'm your AI practice co..."}]
hub-aicoach-chat-1770035435207	What should I practice today?	{"messages": [{"content": "Hi! I'm your AI practice co..."}]
hub-aicoach-chat-1770035366500	What should I practice today?	{"messages": [{"content": "Hi! I'm your AI practice co..."}]
hub-aicoach-chat-1770035155059	Show me my progress	{"messages": [{"content": "Hi! I'm your AI practice co..."}]}

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 a thread quality check. 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, the main title is 'AI Coach Chat Thread Quality Checks' with a timestamp '01/28/26 09:12 PM'. There are filters for 'Threads' (2), 'FretCoach-Hub' (1), and '28/28 (100%)'. A progress bar indicates 'AI Coach Conversation Rating 3.25' and 'User Feedback 0.857142857'. The interface has tabs for 'Queue items' (selected) and 'Configuration'. A search bar allows searching by ID. The main area displays a table of annotation items. Each item includes a checkbox, an ID, the first message content, the last message content, and a comments section with a red error icon and a note about missing details. The table columns are labeled: ID, First message, Last message, and Comments.

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"}]	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"}]	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"}]	Not many details
thread-1769592573338	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.

Datasets:

- Curated test cases from real user sessions
- Used across experiment runs for consistent testing

Prompts:

- Version-controlled coaching prompt templates
- Used in playground for rapid iteration

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	-	-		

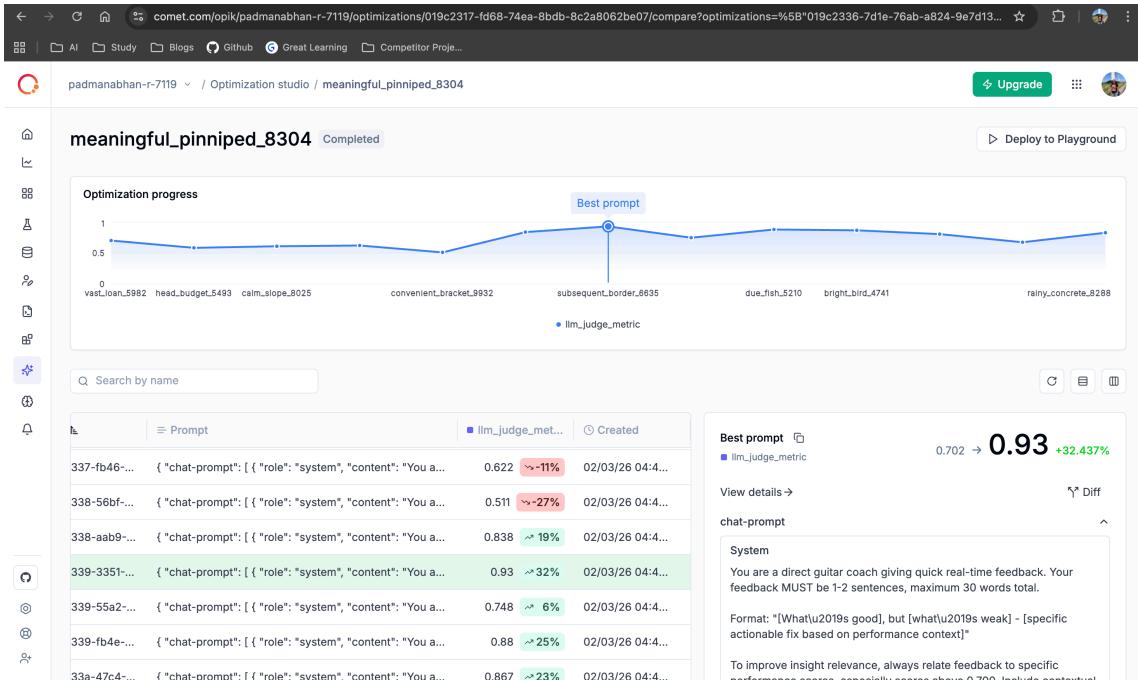
Experiments with custom metric

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 is a sidebar with navigation links like Home, Dashboards, Observability, Projects (which is selected), Evaluation, Experiments, Datasets, Annotation queue, Prompt engineering, Prompt library, Playground, Optimization, and Production. The main area has tabs for 'Trace - 8 spans' and 'Output'. The 'Output' tab displays a 'Pretty' view of a trace with spans for 'agent', 'ChatGoogleGenerativeAI', 'should_continue', 'tools', 'execute_sql_query', and another 'agent' span. A message from 'OpikAssist Beta' suggests focusing on pitch accuracy, timing stability, and scale conformity. It also highlights excessive LLM token usage and long duration for the ChatGoogleGenerativeAI span. A text input field at the bottom allows users to type messages to the AI.

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 are tabs for 'Feedback definitions' (selected), 'AI Providers', 'Workspace preferences', and 'Members'. Below is a search bar and a button to 'Create new feedback definition'. A note says 'Create custom fields to manually rate LLM outputs. Use them to collect structured feedback and track quality over time.' The main area 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: thumbs up, thumbs down).

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.			
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 present. To the right of the search bar are buttons for 'Create new rule' and other actions. The main content is a table with columns: 'Name', 'Projects', 'Scope', 'Status', and 'Logs'. The table lists 11 rules, each with a checkbox, a name, a project list, a scope type (Trace or Thread), an enabled status, a 'Show logs' link, and a three-dot menu. At the bottom of the table, it says 'Showing 1-10 of 11'.

Online evaluation rules dashboard (1-10 of 11)

This screenshot shows the final page of the online evaluation rules dashboard, displaying rules 11 through 11. The interface is identical to the previous screenshot, with a sidebar, a header, and a main table. The table has one visible row for 'hub_user_frustration_score' with the same columns and details as the previous pages. At the bottom, it says 'Showing 11-11 of 11'.

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
--	------------------	---

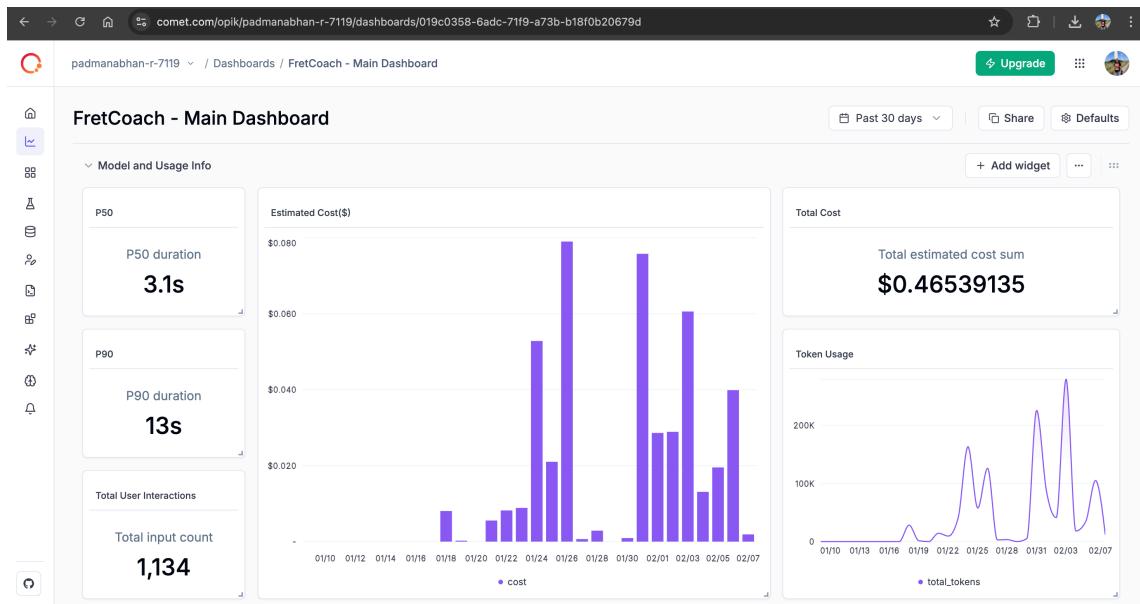
🧠 FretCoach Hub (Web)

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 buttons for 'Upgrade', 'Share', and 'Defaults'. Below the header, a section titled 'FretCoach - AI Evaluation Metrics Info' contains two main sections: 'FretCoach - Studio and Portable (Core Functionality)' and 'FretCoach Hub (Web)'. Each section includes a brief description, a table of metrics with ranges and meanings, and a 'What the Score Means' table.

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

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

AI Evaluation Metrics documentation embedded in dashboard

This screenshot shows the live production dashboard for FretCoach. It features a similar layout to the main dashboard but with more widgets. The top section displays three key metrics: Average Live Coach Feedback Quality (3.432), Average Practice Recommendation Alignment (0.884), and Average Immediate Actionability (0.947). Below these, there are four more sections: Response Clarity (0.972), Context Usage Quality (0.97), Average Hub Actionability (0.994), and Average Hub Data Groundedness (0.834). Each section includes a brief description of the metric and its scale.

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:

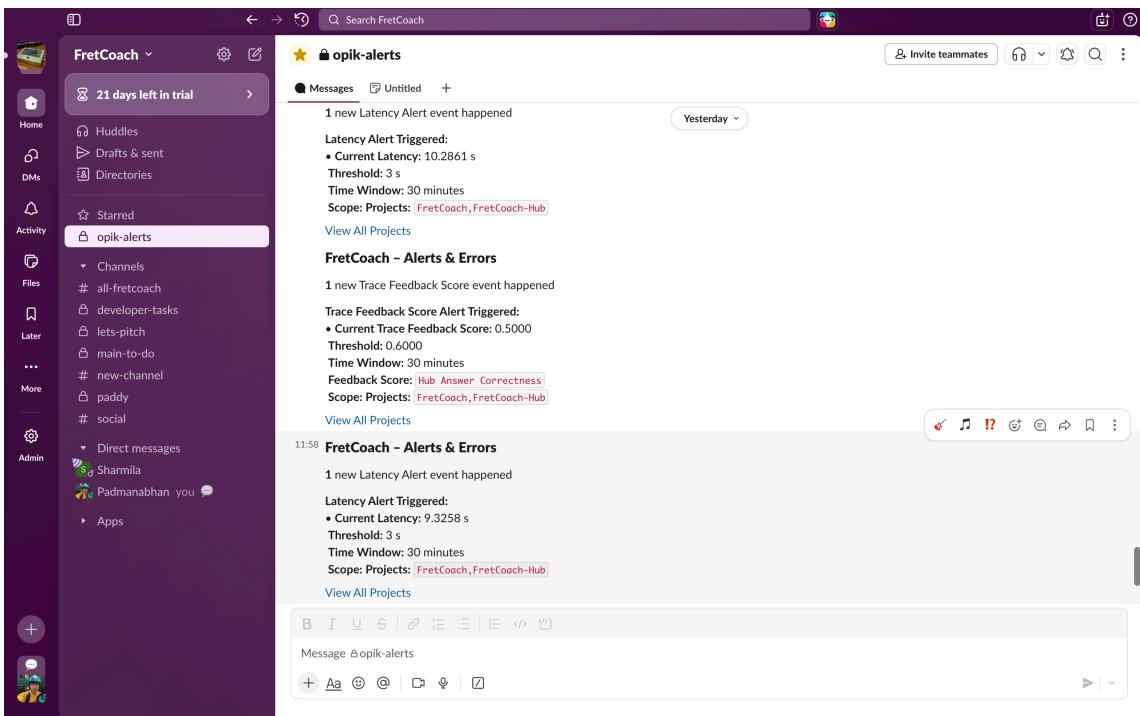
- Proactive issue detection before users report problems
- Real-time visibility into production AI quality
- Team-wide awareness through Slack notifications
- Quick response to quality degradation or system errors

```

1 v {
  "blocks": [
    {
      "type": "header",
      "text": {
        "type": "plain_text",
        "text": "FretCoach - Alerts & Errors"
      }
    },
    {
      "type": "section",
      "text": {
        "type": "mrkdwn",
        "text": "*Trace Errors Alert Triggered*\n*Current Trace Errors*: 15\n*Threshold*: 10\n*Time Window*: 1 hour\n*Scope*: *Projects*: *Demo Project, Default Project*\n*Project*: *\n*Project URL*: http://localhost:5173/demo_workspace_name/projects\n*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 sentence maximum" constraint	Significantly faster responses, more focused feedback
TTS latency spike	TTS taking longer than expected on some calls	Implemented singleton audio player with stop() before new audio	Consistent TTS latency, no audio overlap
Prompt optimization	Live feedback prompt quality measured via llm_judge_metric	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.