

Opik Integration in FretCoach

Features Implemented

1. Traces with Metadata and Tags

All LLM calls are logged as traces in Opik with proper tags to organize and filter different types of interactions.

Hub Coach Chats:

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

The screenshot displays the Opik dashboard interface. On the left, a sidebar menu lists various sections like Home, Dashboards, Observability, Projects, Evaluation, Experiments, Datasets, Annotation queue, Prompt engineering, Prompt library, Playground, Optimization, Optimization studio, Production, Online evaluation, and Alerts. The main area shows a trace for a 'LangGraph' agent. The trace details include a 'ChatGoogleGenerativeAI' call with a duration of 1.6s, a 'should_continue' step, and a 'tools' step with an 'execute_sql_query' action. The right panel shows the agent graph and a table of feedback scores for various metrics.

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 l...
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`
- Used for generating personalized practice recommendations

Trace - 2 spans

- RunnableSequence**
 - ChatOpenAI**

3.1s # 2084 ± 1997/87 <\$0.001 2 5

openai gpt-4o-mini-2024-07-18
 - RunnableLambda**

0.001s

RunnableSequence Details

01/31/26 12:41 PM 3.1s # 2084 <\$0.001 2

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

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
- Used for real-time coaching feedback during practice sessions

Trace - 1 spans

- ChatOpenAI**

2s # 389 ± 361/28 <\$0.001 1 5

openai gpt-4o-mini-2024-07-18

ChatOpenAI Details

01/30/26 10:49 PM 2s # 389 <\$0.001 1

Tags: fretcoach-core, fretcoach-studio, gpt-4o-mini, live-feedback, manual-mode

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 providers:

Live AI feedback traces during practice sessions

2. Thread Management

Each trace uses a structured `thread_id` to group related LLM calls and maintain conversation context across multiple requests.

Thread Naming Conventions:

Hub Coach Chats:

- Format: `hub-{user_id}`
- Example: `hub-default_user`
- Groups all coach chat messages for a user's conversation session

AI Mode (Practice Recommendations):

- Format: `{deployment}-ai-mode-{practice_id}`
- Example: `studio-ai-mode-a1b2c3d4-e5f6-7890-abcd-ef1234567890`
- Maintains thread across multiple recommendation requests until session starts
- Uses `practice_id` to persist thread even when user requests new recommendations

Live AI Feedback:

- Format: `{session_id}-live-aicoach-feedback`
- Example: `abc123-live-aicoach-feedback`
- Groups all live feedback calls within a single practice session

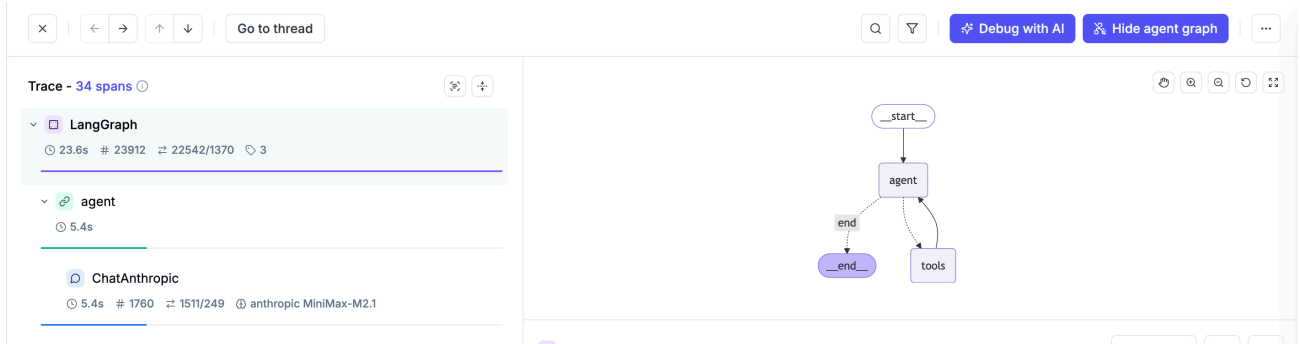
ID	First message	Last message
hub-aicoach-chat-1770250552190	How am I doing compared to average?	{"messages":[{"content":"Hi! I'm your AI practice co...
hub-aicoach-chat-1770223524432	How am I doing compared to average? And what day is ...	{"messages":[{"content":"Hi! I'm your AI practice co...
hub-aicoach-chat-1770217604585	How am I doing compared to average?	{ "messages": [{ "content": "Hi! I'm your AI pract...
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 pract...
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 l...
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 pract...
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 for hub coach chats using `workflow.get_graph(xray=True)`.

Shows 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 to review and annotate agent outputs for quality evaluation.

Implementation:

- Created custom feedback definitions for LLM output quality
- Reviewed agent responses in annotation queues
- Rated outputs using custom criteria
- Human-in-the-loop evaluation for agent improvements

	ID	First message	Last message	Comments
<input type="checkbox"/>	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	Very bad, no practice plan actually displayed.
<input type="checkbox"/>	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
<input type="checkbox"/>	thread-1769592723853	show me the scales that i have showed nost 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
<input type="checkbox"/>	thread-1769592573338	show me the scales that i have showed nost 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	-
<input type="checkbox"/>	thread-1769591135023	Show me my progress	{"messages":[{"content":"Hi! I'm your AI practice coach. I can help you analyze	-

Annotation queue for with reviewed LLM outputs

5. Datasets and Prompts

Created and saved datasets and prompts for use in experiments, evaluations, and playground testing.

Datasets:

- Curated test cases from real user sessions

- Saved to Opik for reproducible evaluations
- Used across experiment runs for consistent testing

Prompts:

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

The screenshot shows the Opik web interface. On the left is a sidebar with navigation options: Dashboards, Observability, Projects (3), Evaluation, Experiments (2), **Datasets (4)**, Annotation queues (2), Prompt engineering, Prompt library (1), Playground, Optimization, and Optimization studio (2). The main content area is titled 'padmanabhan-r-7119 / Datasets'. It includes a search bar and a 'Create new dataset' button. Below is a table of datasets:

	Name	Description	# Item co...	Most recent exper...	Created	
<input type="checkbox"/>	fretCoach_live_ai_feedback_val_9		9	02/03/26 05:06 PM	02/03/26 04:11 PM	...
<input type="checkbox"/>	fretcoach_live_ai_feedback_train_25		25	02/03/26 04:48 PM	02/03/26 04:11 PM	...
<input type="checkbox"/>	FretCoach Live AI Feedback		44		02/03/26 03:48 PM	...
<input type="checkbox"/>	FretCoach Hub AI Coach Chat		10		01/27/26 10:24 AM	...

At the bottom right of the table, it says 'Showing 1-4 of 4'.

Datasets created for experiment runs

The screenshot shows the Opik web interface for a specific prompt. The breadcrumb is 'padmanabhan-r-7119 / Prompt library / Live AI Feedback - System Prompt'. The prompt title is 'Live AI Feedback - System Prompt' with a timestamp '02/03/26 05:09 PM'. There are tabs for 'Prompt', 'Experiments', and 'Commits'. Below the tabs are buttons: 'Use this prompt', 'Try in the Playground', 'Improve prompt', and 'Edit prompt'. The main area shows the prompt text:

You are a direct guitar coach giving quick real-time feedback. Your feedback MUST be 1-2 sentences, maximum 30 words total.

Format: "[What's good], but [what's weak] - [specific actionable fix based on performance context]"

To improve insight relevance, always relate feedback to specific performance scores, especially scores above 0.700. Include contextual details from the player's playing style to inform suggestions:

- Pitch Accuracy: How cleanly notes are fretted (low = finger pressure issues)
→ Fix: "ease finger pressure to improve note clarity" or "focus on clean fretting by adjusting finger placement"
- Scale Conformity: Playing correct scale notes across fretboard positions (low = stuck in one position or wrong notes)
→ Fix: "explore positions 5-7 to enhance versatility" or "move up the fretboard to discover new notes"
- Timing Stability: Consistency of note spacing (low = rushing, dragging, uneven rhythm)
→ Fix: "use a metronome at 60 BPM to develop timing" or "slow down and count to create consistent spacing"

Be direct and conversational, and vary your wording. Ensure your suggestions are anchored in the player's specific

On the right, there is a 'Commit history' section showing a single commit with hash '4b4635b6' at '02/03/26 05:09 PM'.

Saved prompts

6. Experiments and Custom Metrics

Ran experiments to evaluate LLM performance using both default Opik metrics and custom-created metrics.

Default Metrics:

- Used Opik's built-in evaluation metrics
- Measured response quality across test datasets

Custom Metrics:

- Created custom metric for domain-specific evaluation
- Tailored to guitar coaching context
- Measured 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

Upgrade

fretcoach-default-metrics-eval

Details Dashboards

02/03/26 05:05 PM fretCoach_live_ai_feedback_val_9 Traces

answer_relevance_metric (avg) 0.904444444 context_precision_metric (avg) 0.788888889 context_recall_metric (avg) 0.838888889 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 expected_out...	input	Evaluation task context	input	output	reference	Duration p50 67.6s	# Total token avg 13042.667
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	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
019c2325-43c2-73b...	Timing is decent	Enabled metrics:...	0: "You are a direct guitar coach giving	Enabled metrics:...	Good timing etah	Timing is decent	67.8s	12,82

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

Upgrade

fretcoach-coaching-feedback-eval

02/03/26 05:01 PM fretCoach_live_ai_feedback_val_9 Traces

coaching_quality (avg) 0.788888889

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 expected_out...	input	Evaluation task input	output	reference	Duration p50 3.1s	# Total tokens p50 0	Estimated co avg 0
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 accur...	Your pitch accur...	4.1s	-	
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	-	

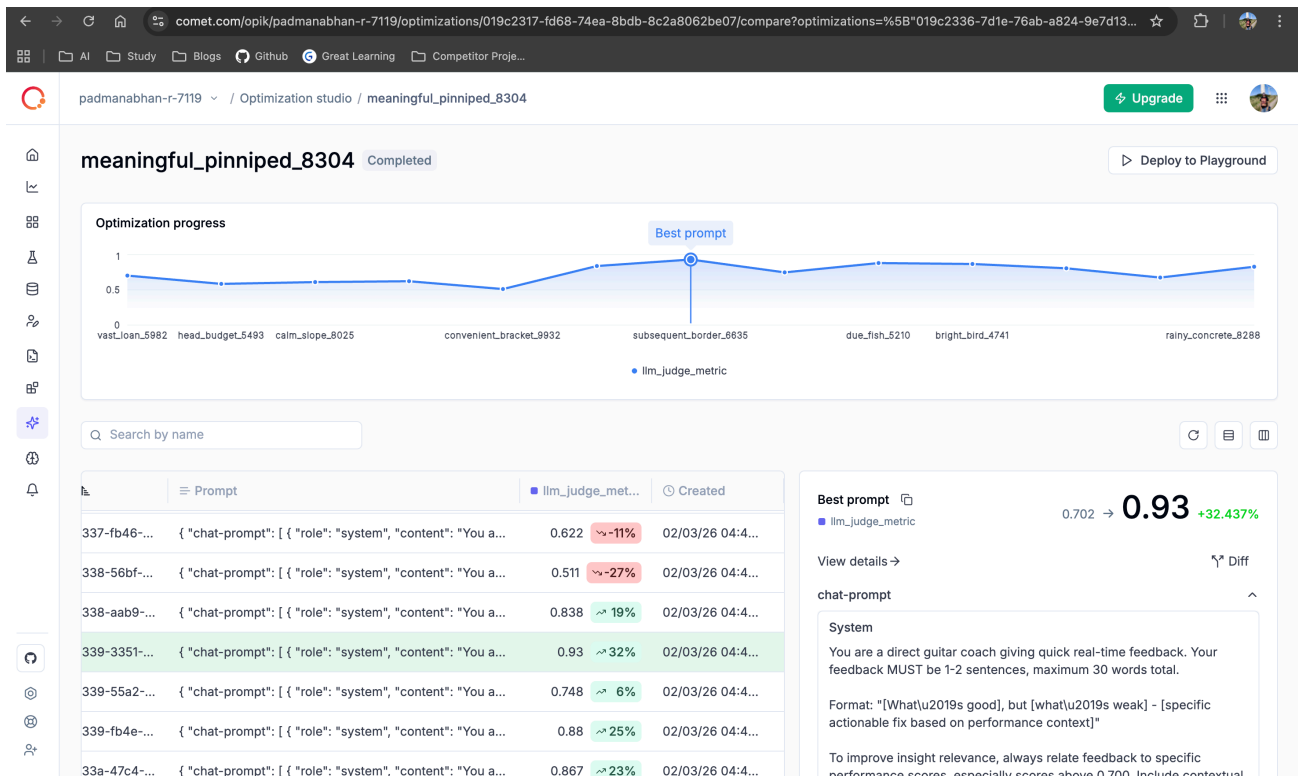
Experiments with custom metric

7. Optimization Studio

Used Optimization Studio to improve the prompt used in the live feedback module.

Results:

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



Optimization Studio results for live feedback prompt