Other Features of LangSmith

Advanced Capabilities for LLM Observability, Evaluation, and Collaboration

1 Monitoring and Alerting

What it Does

Monitoring in LangSmith looks across multiple traces to assess the overall health of your LLM system. It aggregates key metrics like:

- **Latency:** P50, P95, P99 percentiles
- **Token Usage and Cost**
- **Error and Success Rates**

You can set up **alerts** to notify you when metrics drift outside acceptable ranges, such as a sudden increase in latency or cost spikes.

Why it Matters

Issues in production often appear as patterns across runs rather than in a single trace. Monitoring helps detect early warning signals before they affect users — allowing for proactive action rather than reactive troubleshooting.

2 Evaluation

What it Does

Evaluation in LangSmith systematically measures the quality of your LLM outputs using:

- Gold-standard datasets
- Custom evaluation metrics (faithfulness, relevance, completeness)
- LLM-as-a-judge or semantic similarity scoring
- Custom Python evaluators

Evaluations can be executed both:

- **Offline:** Batch testing before deployment
- **Online:** Continuous evaluation on live traffic

Why it Matters

LLM behavior can vary unpredictably with minor changes in prompts or models. Evaluation ensures improvements are consistent, repeatable, and prevent regressions.

Example: For a RAG chatbot:

- **Faithfulness:** Are responses grounded in retrieved documents?
- **Relevance:** Does the answer address the user's query?

You can benchmark performance across multiple models (e.g., GPT-4, Claude, LLAMA) to select the best configuration.

3 Prompt Experimentation

What it Does

LangSmith enables systematic testing and comparison of prompt variations through:

- **A/B Testing:** Run multiple prompts on identical datasets
- **Performance Tracking:** Compare results against evaluation metrics
- **Version History:** Record outcomes to track which prompt performs best over time

Why it Matters

Prompt design greatly influences model performance. Experimentation helps refine prompts systematically instead of relying on intuition.

4 Dataset Creation & Annotation

What it Does

Provides powerful tools to:

- Build high-quality datasets for evaluation or fine-tuning
- Manually annotate responses (e.g., correctness labels)
- Maintain versioned datasets for reuse across projects

Why it Matters

High-quality, curated datasets form the foundation for evaluation and iterative feedback loops.

Example: For a customer support assistant, build a dataset of frequent questions and correct answers. Use this dataset to benchmark your RAG agent each time retrieval logic changes.

5 User Feedback Integration

What it Does

Captures real-world user reactions:

- Thumbs up/down or star ratings
- Structured or free-form feedback

Feedback is stored with the corresponding trace, prompt, and model version — enabling bulk analysis of user satisfaction.

Why it Matters

Direct feedback helps prioritize improvements and aligns system behavior with actual user preferences.

6 Collaboration

What it Does

LangSmith supports team-based collaboration by allowing:

- Shared access to traces, datasets, and evaluations
- Web UI for PMs, QA, and annotators to inspect or comment
- Shared dashboards for monitoring experiments

Why it Matters

Encourages cross-functional visibility and smooth collaboration between engineers, data scientists, and non-technical stakeholders.

LangSmith brings observability, evaluation, and collaboration together — empowering teams to build trustworthy, data-driven LLM applications.