Production: Batch, Optimization, and Observability

Course syllabus

Version 2026-11

Overview

Move to production: batch inference up to 1M requests, predicted outputs, citations with sources, and observability (traces, judges, offline evaluations).

Learning objectives

- Prepare a batch inference job in JSONL, track its status, and retrieve the results
- Reduce response latency with predicted outputs on rewriting tasks
- Implement citations with reference object to ground responses in their sources
- Trace an application with OpenTelemetry, explore production traces, and automate evaluation with LLM judges
- Build datasets from traces and measure each version with offline evaluations

Audience and prerequisites

Developers, engineers, and integrators who want to use Mistral AI via API in their applications.

Proven experience developing applications and deploying them to production.

Duration and format

Estimated duration : 1 h 15 min. 100% online, self-paced course, accessible 24/7 from a browser. Progress is saved, and you can resume at any time. Estimated duration: 1 h 15 min.

Content

- Batch Inference (4 lessons)
- Predicted Outputs (3 lessons)
- Citations and References (3 lessons)
- Observability (6 lessons)

Assessment

Self-assessment based on the Key takeaways that conclude each lesson, and direct practice of the examples on your own examples. A final quiz of 5 graded questions at the end of the last lesson validates what you have learned.

Accessibility

Course accessible to people with disabilities: text content compatible with screen readers, keyboard navigation, and compliant color contrast. For any adaptation needs, email hello@corsen.ai.

Price and recognition

Free

Certificate of completion issued by Corsen AI.

Access time

Immediate after account creation and email confirmation.

Corsen AI — Paris, France — SIRET 10297040700015
academy@corsen.ai — https://academy.corsen.ai