CORSEN AI Academy

Production, Evals, and Optimization

Course syllabus

Version 2026-11

Overview

Covers latency, prompt caching, Batch API, evals, compaction, and complete production architecture.

Learning objectives

- Measure end-to-end latency of a call and identify priority optimization levers
- Reduce costs at scale with prompt caching, Predicted Outputs, and the Batch API
- Design an asynchronous architecture based on queues, parallelism, and background mode
- Master long conversations through compaction, token counting, and budgeting
- Automate evaluations with the Evals framework and integrate them into a CI/CD pipeline

Audience and prerequisites

Developers, engineers, and integrators looking to leverage ChatGPT via API in their applications.

Confirmed experience in application development and deployment to production.

Duration and format

Estimated duration : 2 h. 100% online self-paced course, accessible 24/7 from a browser. Progress is saved, allowing you to resume at any time. Estimated duration: 2 h.

Content

- Optimizing performance (5 lessons)
- Scale Processing (1 lessons)
- Scaling Processing (4 lessons)
- Context management (4 lessons)
- Evaluation (4 lessons)
- Complete production (2 lessons)

Assessment

Self-assessment based on the Key takeaways concluding each lesson, and direct implementation of examples on your own examples. A final quiz of 5 corrected questions at the end of the last lesson validates learning.

Accessibility

Self-paced course accessible to people with disabilities: text content compatible with screen readers, keyboard navigation, and compliant contrasts. 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