

Ada Lovelace

Senior Software Engineer

 ada@example.com  +1 (555) 123-4567  San Francisco, CA  GitHub  LinkedIn

Summary

Senior engineer with 8+ years building distributed systems and developer platforms. Led the red of a payments pipeline processing \$2B annually, cutting latency 40% and cost 30%. Passionate ab systems that scale cleanly and teams that ship calmly.



Experience

Senior Software Engineer at Anthropic (Tenure 1)2023 - Present
San Francisco

- CROSSPAGE-X9K2 Led a multi-quarter initiative to redesign the request pipeline from ingr through the model router, eliminating three legacy proxies, consolidating four telemetry sy into one, and reducing end-to-end p99 latency from 240ms to 78ms while cutting monthly infrastructure spend by \$340K and improving developer velocity measurably across the org
- Led the design of a streaming inference pipeline serving 1M+ daily requests at p99 < 80ms
- Cut inference cost 38% by re-architecting the model router on a tiered GPU fleet.
- Mentored 4 engineers; established the team's on-call runbooks and incident review proces

Senior Software Engineer at Anthropic (Tenure 2)2023 - Present
San Francisco

- CROSSPAGE-X9K2 Led a multi-quarter initiative to redesign the request pipeline from ingr through the model router, eliminating three legacy proxies, consolidating four telemetry sy into one, and reducing end-to-end p99 latency from 240ms to 78ms while cutting monthly infrastructure spend by \$340K and improving developer velocity measurably across the org
- Led the design of a streaming inference pipeline serving 1M+ daily requests at p99 < 80ms
- Cut inference cost 38% by re-architecting the model router on a tiered GPU fleet.
- Mentored 4 engineers; established the team's on-call runbooks and incident review proces

Senior Software Engineer at Anthropic (Tenure 3)2023 - Present
San Francisco

- CROSSPAGE-X9K2 Led a multi-quarter initiative to redesign the request pipeline from ingr through the model router, eliminating three legacy proxies, consolidating four telemetry sy into one, and reducing end-to-end p99 latency from 240ms to 78ms while cutting monthly infrastructure spend by \$340K and improving developer velocity measurably across the org
- Led the design of a streaming inference pipeline serving 1M+ daily requests at p99 < 80ms
- Cut inference cost 38% by re-architecting the model router on a tiered GPU fleet.

- • Mentored 4 engineers; established the team's on-call runbooks and incident review process.

Senior Software Engineer at Anthropic (Tenure 4)2023 - Present

San Francisco

- • CROSSPAGE-X9K2 Led a multi-quarter initiative to redesign the request pipeline from ingress through the model router, eliminating three legacy proxies, consolidating four telemetry systems into one, and reducing end-to-end p99 latency from 240ms to 78ms while cutting monthly infrastructure spend by \$340K and improving developer velocity measurably across the org.
- • Led the design of a streaming inference pipeline serving 1M+ daily requests at p99 < 80ms.
- • Cut inference cost 38% by re-architecting the model router on a tiered GPU fleet.
- • Mentored 4 engineers; established the team's on-call runbooks and incident review process.

Senior Software Engineer at Anthropic (Tenure 5)2023 - Present

San Francisco

- • CROSSPAGE-X9K2 Led a multi-quarter initiative to redesign the request pipeline from ingress through the model router, eliminating three legacy proxies, consolidating four telemetry systems into one, and reducing end-to-end p99 latency from 240ms to 78ms while cutting monthly infrastructure spend by \$340K and improving developer velocity measurably across the org.
- • Led the design of a streaming inference pipeline serving 1M+ daily requests at p99 < 80ms.
- • Cut inference cost 38% by re-architecting the model router on a tiered GPU fleet.
- • Mentored 4 engineers; established the team's on-call runbooks and incident review process.

Senior Software Engineer at Anthropic (Tenure 6)2023 - Present

San Francisco

- • CROSSPAGE-X9K2 Led a multi-quarter initiative to redesign the request pipeline from ingress through the model router, eliminating three legacy proxies, consolidating four telemetry systems into one, and reducing end-to-end p99 latency from 240ms to 78ms while cutting monthly infrastructure spend by \$340K and improving developer velocity measurably across the org.
- • Led the design of a streaming inference pipeline serving 1M+ daily requests at p99 < 80ms.
- • Cut inference cost 38% by re-architecting the model router on a tiered GPU fleet.
- • Mentored 4 engineers; established the team's on-call runbooks and incident review process.

<>

Projects

OpenMetrics at Open-source Prometheus-compatible metrics aggregator.2022 - Present

- • 1.2k GitHub stars; adopted by 3 mid-size companies.

OpenMetrics at Open-source Prometheus-compatible metrics aggregator.2022 - Present

- • 1.2k GitHub stars; adopted by 3 mid-size companies.

OpenMetrics at Open-source Prometheus-compatible metrics aggregator.2022 - Present

- • 1.2k GitHub stars; adopted by 3 mid-size companies.



Skills

Languages:TypeScriptGoPythonRust
Infrastructure:PostgresKubernetesTerraformRedis
Domains:Distributed SystemsLLM InfraPayments
Frontend:ReactNext.jsTailwindVite
Cloud:AWSGCPCloudflareFly.io



Education

B.S. Computer Science at Stanford University2012 - 2016
Systems & AI
B.S. Computer Science at Stanford University2012 - 2016
Systems & AI



Certifications

AWS Solutions Architect — Professional
Certified Kubernetes Administrator (CKA)
Google Cloud Professional Cloud Architect



Languages

EnglishNative
MandarinProfessional