



UNIVERSIDAD AUTÓNOMA DE CHIAPAS
Escuela de Tecnologías Digitales Aplicadas
Licenciatura en Inteligencia Artificial y Ciencia de Datos



Unidad de Competencia							
Arquitecturas de IA a Gran Escala							
Cuatrimestre	Fecha de elaboración			Modalidad	Tipo	Área de formación	
9	DD	MM	AÑO	• Escolarizada – Presencial / Laboratorio especializado	• Curso–Laboratorio	Programación e ingeniería de software	
	27	11	2025				
Perfil docente							
• Formación: ◦ Ingeniería en IA, Ciencia de Datos, Ingeniería de Software, Computación, Sistemas Distribuidos o áreas afines. ◦ Posgrado deseable en Deep Learning, Sistemas Distribuidos, Arquitecturas de IA o Cloud. • Experiencia profesional: ◦ Diseñando sistemas de IA en producción. ◦ Dominio de arquitecturas distribuidas, GPUs, clusters, MLOps avanzado, servidores de modelos y optimización a gran escala. ◦ Experiencia práctica con frameworks como: Ray, Horovod, DeepSpeed; Kubernetes, Kubeflow, MLflow; vLLM, Triton Inference Server; Infraestructura cloud (AWS/GCP/Azure) • Dominio de inglés: ◦ B2–C1 (lectura de papers de arquitectura y escalabilidad). • Competencias digitales: ◦ Python avanzado, GPU computing, APIs, Docker/Kubernetes, despliegue distribuido.							
Horas con Mediación Docente Cuatrimestrales		Horas con Mediación Docente semana/mes		Horas de Actividad Independiente	Horas de Práctica de Campo	Créditos	Academia
HT	HP	HT	HP	48	0	7	
32	32	2	2				
Horas de docencia total de la unidad de competencia: 64							



UNIVERSIDAD AUTÓNOMA DE CHIAPAS
Escuela de Tecnologías Digitales Aplicadas
Licenciatura en Inteligencia Artificial y Ciencia de Datos



Propósito general:	Presentación:
El estudiantado diseña, implementa y opera arquitecturas de inteligencia artificial a gran escala mediante técnicas de paralelización, distribución de modelos y datos, optimización de inferencia, uso de clústeres GPU/CPU y despliegue empresarial, garantizando eficiencia, seguridad, rendimiento y escalabilidad.	La inteligencia artificial moderna exige arquitecturas masivas capaces de entrenar y desplegar modelos con millones o billones de parámetros. Esta asignatura ofrece una visión integral de: entrenamiento distribuido, paralelización de modelos y datos, optimización para inferencia a gran escala, infraestructura GPU/TPU, servidores de modelos, escalamiento horizontal y vertical, arquitecturas para LLMs y deep learning avanzado. Se articula con: IA Generativa y LLMs, MLOps, Ingeniería de Datos a Gran Escala, Sistemas Ciberfísicos, Auditoría Algorítmica y Gobernanza.
Competencias	
Genéricas	Disciplinares
• Pensamiento computacional avanzado. • Resolución de problemas complejos. • Comunicación técnica. • Trabajo colaborativo.	• Diseña arquitecturas distribuidas de IA. • Emplea paralelización y optimización de modelos. • Automatiza y despliega modelos en clusters.
Profesionales	
• Opera infraestructuras de IA reales. • Evalúa rendimiento, costos, riesgos y eficiencia. • Implementa soluciones de IA escalables para organizaciones.	
Producto final	
Arquitectura de IA a Gran Escala Funcional, que incluya: Diseño de arquitectura distribuida, Pipeline de entrenamiento o inferencia masiva, Uso de clúster GPU o simulación, Servidor de modelo optimizado (Triton o vLLM), Dashboard de monitoreo y Documento técnico + justificación de escalabilidad	
Enseñanza con perspectiva de género, interculturalidad y Derechos Humanos	
Con la intención de propiciar una enseñanza con perspectiva de género, interculturalidad y derechos humanos, la o el docente que imparta esta unidad de competencia deberá: • Eliminar y rechazar actitudes y comportamientos sexistas, clasistas, racistas y todo tipo de estereotipos que supongan discriminación y	



UNIVERSIDAD AUTÓNOMA DE CHIAPAS
Escuela de Tecnologías Digitales Aplicadas
Licenciatura en Inteligencia Artificial y Ciencia de Datos



violencia en los ámbitos de enseñanza-aprendizaje (virtual y presencial), a partir del ejemplo en su práctica docente.

- Promover que el desarrollo del programa educativo y material a utilizar se base en el respeto, protección, promoción y garantía de los derechos humanos, de la dignidad e integridad de todas las personas.
- Promover la eliminación de estereotipos o cualquier otra práctica de discriminación o denigración de la persona, que pueda presentarse mediante el uso del lenguaje, el empleo de ejemplos, la diferenciación entre grupos, origen étnico o cualquier otra circunstancia.
- Respetar la identidad de género de las personas en todo momento y nombrarlas tal y como ellas lo deseen.
- Respetar y promover la igualdad de trato y de oportunidades entre las y los estudiantes, favoreciendo acciones afirmativas en beneficio de los grupos en situación de vulnerabilidad, incluso en las afectaciones personales provocadas por la pandemia COVID-19
- Promover y usar lenguaje incluyente y no sexista tanto en el salón de clases como en los trabajos, tesis y otros documentos que elabore la comunidad estudiantil.

Estos elementos son motivados por la información asequible en los programas de estudio de la Licenciatura en Derecho, Modalidad Educativa Presencial (sistema escolarizado), Tomo II, de la Universidad Nacional Autónoma de México (UNAM 2020).



UNIVERSIDAD AUTÓNOMA DE CHIAPAS
Escuela de Tecnologías Digitales Aplicadas
Licenciatura en Inteligencia Artificial y Ciencia de Datos



Nivel Unidad/Subcompetencias	
Nombre de la Subcompetencia 1	Elementos de la subcompetencia 1
Entrenamiento Distribuido y Paralelización de Modelos	Conocimientos: • Paralelización: Data Parallelism, Model Parallelism, Pipeline Parallelism • Frameworks: Horovod, DeepSpeed, Distributed PyTorch • Uso de clusters GPU/TPU • Comunicación: NCCL, gRPC Habilidades: • Uso de entrenamiento distribuido • Optimización de recursos • Manejo de contenedores GPU Valores y actitudes: • Responsabilidad técnica • Rigor computacional
Número de semanas	
4 semanas	
Propósito de la subcompetencia	
El estudiantado implementa estrategias de paralelización para entrenar modelos a gran escala.	
Estrategias Didácticas	Recursos didácticos
<i>Las estrategias y recursos didácticos consideran la inclusión y perspectiva de género, así como la diversidad cultural.</i>	
• Laboratorio de paralelización • Visualización del DAG de entrenamiento • Mini-pipeline distribuido	• PyTorch Distributed • Horovod • Ray Train

Nombre de la Subcompetencia 2	Elementos de la subcompetencia 2
Servidores de Modelos e Inferencia a Gran Escala	Conocimientos: • Triton Inference Server
Número de semanas	



UNIVERSIDAD AUTÓNOMA DE CHIAPAS
Escuela de Tecnologías Digitales Aplicadas
Licenciatura en Inteligencia Artificial y Ciencia de Datos



4 semanas	• vLLM para LLMs • Inferencia optimizada (batching, caching, quantization) • Distribución de cargas y autoscaling • Métricas clave: throughput, latencia, QPS Habilidades: • Configurar servidores de modelos • Optimizar latencia y rendimiento • Escalar inferencia Valores y actitudes: • Responsabilidad operativa • Ética en despliegue
Propósito de la subcompetencia	
El estudiantado despliega modelos optimizados y mide su rendimiento a escala.	
Estrategias Didácticas	Recursos didácticos
<i>Las estrategias y recursos didácticos consideran la inclusión y perspectiva de género, así como la diversidad cultural.</i>	
• Laboratorio Triton/vLLM • Pruebas de rendimiento • Comparación de estrategias	• Triton Server • vLLM • Docker + GPU

Nombre de la Subcompetencia 3	Elementos de la subcompetencia 3
Orquestación, Automatización y Cloud para IA Masiva	Conocimientos: • Kubernetes (K8s) para IA • Kubeflow, MLflow, Ray Serve • Infraestructura como código (IaC) • Monitoreo con Prometheus + Grafana • Gestión de costos Habilidades:
Número de semanas	
4 semanas	
Propósito de la subcompetencia	
El estudiantado opera clústeres de IA y automatiza tareas mediante pipelines y orquestadores.	



UNIVERSIDAD AUTÓNOMA DE CHIAPAS
Escuela de Tecnologías Digitales Aplicadas
Licenciatura en Inteligencia Artificial y Ciencia de Datos



	• Deploy en Kubernetes • Orquestación de modelos • Monitoreo continuo
	Valores y actitudes: • Orden técnico • Eficiencia en recursos
Estrategias Didácticas	Recursos didácticos
<i>Las estrategias y recursos didácticos consideran la inclusión y perspectiva de género, así como la diversidad cultural.</i>	
• Taller de K8s • Creación de pipeline automatizado • Dashboard de monitoreo	• Kubernetes • Kubeflow • Prometheus/Grafana

Nombre de la Subcompetencia 4	Elementos de la subcompetencia 4
Arquitecturas y Sistemas para LLMs y Deep Learning Masivo	Conocimientos: • Arquitecturas para LLMs: sharding, serving, embeddings a escala • Vector databases distribuidas • Retrieval a escala • Topologías de alta disponibilidad • Seguridad y gobernanza del modelo Habilidades: • Diseño arquitectónico • Integración avanzada de LLM pipelines • Evaluación técnica–ética de despliegues masivos Valores y actitudes: • Precisión
Número de semanas	
4 semanas	
Propósito de la subcompetencia	
El estudiantado diseña arquitecturas especializadas para aplicaciones generativas y modelos avanzados de IA.	



UNIVERSIDAD AUTÓNOMA DE CHIAPAS
Escuela de Tecnologías Digitales Aplicadas
Licenciatura en Inteligencia Artificial y Ciencia de Datos



	• Responsabilidad social
Estrategias Didácticas	Recursos didácticos
<i>Las estrategias y recursos didácticos consideran la inclusión y perspectiva de género, así como la diversidad cultural.</i>	
• Propuesta arquitectónica • Simulación de carga • Justificación técnica–ética	• LangChain • FAISS / Milvus • vLLM + Serving pipelines

Evaluación	
La evaluación de los aprendizajes se desarrollará de forma continua durante el proceso de enseñanza-aprendizaje a través de los siguientes momentos: Evaluación diagnóstica: Evaluación de conocimientos previos de MLOps, DL y cloud. Evaluación formativa: Laboratorios progresivos, Experimentos distribuidos, Reportes de rendimiento y Retroalimentación continua Evaluación sumativa: Arquitectura final de IA, Examen técnico, Presentación profesional y Portafolio digital	
Estrategia de evaluación de los aprendizajes:	Instrumentos de Evaluación:
Estrategias que el estudiante desarrolla para la elaboración de las evidencias de aprendizaje. • Código y configuración • Pipeline distribuido • Servidor de inferencia • Informe técnico	• Rúbricas • Listas de cotejo • Pruebas de rendimiento • Revisión de arquitectura de la IA a gran escala
Referencias	
<i>Las referencias consideran la inclusión y perspectiva de género, así como la diversidad cultural.</i>	
Básicas:	Complementarias:
• DeepSpeed Documentation	• Papers sobre entrenamiento distribuido



UNIVERSIDAD AUTÓNOMA DE CHIAPAS
Escuela de Tecnologías Digitales Aplicadas
Licenciatura en Inteligencia Artificial y Ciencia de Datos



- Ray Documentation
- Triton Inference Server Docs
- vLLM Documentation
- Kubernetes Official Documentation

- AWS/GCP/Azure LLM architecture guides
- ML Systems for Scalability (Stanford CS329S)

Videos:

Bases de Datos:

- CONRICYT
- ArXiv (ML Systems & Scalability)
- Repositorios cloud oficiales