AGENTES DE IA: RUTA ESENCIAL

CÓDIGO	AUTOR	DURACIÓN ESTIMADA	NIVEL DE FORMACIÓN
01B06C02	Carlos Ojeda Sánchez	50 h	Avanzado

Dirigido a

Profesionales de cualquier sector que quieran adquirir una visión práctica del diseño seguro de agentes de IA mediante una alternativa más breve al curso completo, especialmente aplicable a las necesidades de las pymes.

Descripción

Con este contenido de curso profesional el alumnado aprenderá los fundamentos necesarios para identificar, especificar y diseñar agentes de IA aplicables a situaciones profesionales. Se priorizarán la selección de casos de uso, la definición del nivel de autonomía, el diseño de flujos y del system prompt, así como la incorporación de guardrails, medidas de seguridad y mecanismos de intervención humana.

COMPETENCIAS

1. Analizar la arquitectura de un agente de IA, diferenciando sus componentes y ciclo operativo para valorar su aplicación profesional.
2. Clasificar niveles de autonomía en sistemas de IA según supervisión, riesgo y capacidad de decisión para seleccionar usos adecuados.
3. Identificar patrones agénticos en procesos empresariales, evaluando valor, complejidad y riesgo para priorizar casos de uso viables.
4. Especificar un agente mediante un documento estructurado de requisitos para orientar su diseño, construcción, prueba y supervisión.
5. Diseñar flujos agénticos con pasos, decisiones y ramificaciones para representar procesos ejecutables y controlables.
6. Redactar system prompts para agentes, estructurando rol, instrucciones, límites y salidas para guiar comportamientos consistentes.
7. Implementar guardrails de seguridad en agentes, diseñando límites, validaciones y respuestas controladas para reducir riesgos operativos.
8. Aplicar medidas de protección de datos y contención de fallos en agentes para reducir exposición, errores y usos indebidos.
9. Diseñar protocolos de escalado humano, definiendo patrones y puntos de intervención para controlar decisiones sensibles o inciertas.

CRITERIOS DE EVALUACIÓN (Objetivos)

1. Definir los componentes principales de un agente de IA y explicar cómo interactúan durante el ciclo operativo. Diferenciar percepción, razonamiento, memoria y acción mediante ejemplos vinculados a procesos empresariales habituales. Representar el ciclo agéntico completo identificando entradas, decisiones, acciones, retroalimentación y posibles puntos de control. Valorar limitaciones y dependencias de la arquitectura agéntica para evitar interpretaciones

simplificadas o excesivamente automatistas.

2. Describir los niveles de autonomía agéntica y diferenciarlos mediante características observables y ejemplos de uso. Clasificar casos empresariales según su grado de autonomía, justificando la decisión con criterios de supervisión y riesgo. Identificar ventajas, límites y condiciones de uso de cada nivel para evitar delegaciones inadecuadas. Proponer el nivel de autonomía más adecuado para un proceso concreto, considerando impacto, control humano y complejidad operativa.
3. Describir los principales patrones agénticos y asociarlos con necesidades profesionales o procesos empresariales concretos. Evaluar procesos propios para detectar oportunidades de automatización agéntica con valor operativo y bajo riesgo. Priorizar casos de uso considerando impacto esperado, dificultad técnica, datos disponibles y necesidad de supervisión. Reconocer patrones poco adecuados cuando el proceso exige juicio humano, alta sensibilidad o control normativo estricto.
4. Identificar las secciones necesarias de una especificación de agente y explicar la función de cada una. Redactar objetivos, límites, entradas, salidas y capacidades de un agente manteniendo coherencia con el caso de uso. Definir restricciones, riesgos y criterios de aceptación que faciliten la prueba y validación posterior del agente. Revisar una especificación agéntica detectando ambigüedades, omisiones o decisiones de diseño insuficientemente justificadas.
5. Representar flujos de trabajo agénticos utilizando pasos, decisiones, condiciones, entradas y salidas de forma ordenada. Diseñar un flujo multi-paso coherente con un objetivo profesional, incorporando validaciones y puntos de control. Identificar bucles, excepciones y dependencias entre tareas para evitar secuencias ambiguas o difíciles de automatizar. Justificar la estructura del flujo relacionando cada fase con el resultado esperado y el nivel de supervisión necesario.
6. Estructurar un system prompt con identidad, objetivo, contexto, instrucciones, restricciones, formato de salida y criterios de actuación. Redactar instrucciones claras y verificables que reduzcan ambigüedades en la ejecución del agente. Incorporar límites, excepciones y pautas de escalado para controlar respuestas inadecuadas o fuera de alcance. Evaluar un system prompt detectando contradicciones, instrucciones insuficientes o riesgos de comportamiento no deseado.
7. Diferenciar tipos de guardrails y explicar su función preventiva, correctiva o limitadora en sistemas agénticos. Diseñar reglas de validación y bloqueo para controlar entradas, acciones, salidas y uso de herramientas sensibles. Aplicar criterios de seguridad y proporcionalidad para seleccionar guardrails adecuados al riesgo del caso de uso.
8. Especificar medidas básicas de protección de datos aplicables al diseño y uso de agentes en entornos profesionales. Explicar el riesgo de prompt injection y proponer medidas de mitigación proporcionadas al contexto de uso.
9. Diferenciar patrones de escalado humano y relacionarlos con niveles de riesgo, incertidumbre o impacto. Diseñar un protocolo de escalado completo que indique activadores, responsables, tiempos, evidencias y acciones posteriores.

CONTENIDOS

Unidad 1. Anatomía de un agente de IA

1. Definición formal de agente: percibe, razona, actúa
2. 4 componentes: percepción, razonamiento, acción, memoria
3. El ciclo agéntico: percibir → razonar → actuar → actualizar
4. Agente reactivo vs deliberativo
5. El LLM como cerebro: capacidades y limitaciones
6. Diagramas de arquitectura agéntica

Unidad 2. El espectro de la autonomía

1. El espectro: no binario, gradual

2. Nivel 1 Asistente; Nivel 2 Copiloto; Nivel 3 Agente supervisado; Nivel 4 Agente autónomo acotado; Nivel 5 Agente autónomo amplio
3. Criterios: riesgo, reversibilidad, frecuencia, coste del error
4. La autonomía como dial, no como interruptor

Unidad 3. Casos de uso y patrones agénticos

1. 6 patrones: investigación autónoma, procesamiento de documentos, atención y soporte, coordinación y scheduling, monitorización y alerta, creación y publicación
2. Antipatrones: cuándo NO usar un agente
3. Marco: ¿es candidato a agente?

Unidad 4. Especificación de un agente / DEA

1. Por qué especificar antes de construir
2. El DEA (7 secciones): identidad, objetivos, capacidades, restricciones y guardrails, inputs/outputs, criterios de éxito, interacciones
3. Plantilla DEA completa; buenos vs malos ejemplos; checklist de calidad

Unidad 5. Diseño de flujos multi-paso

1. Pensar en flujos: la mente procesal
2. Elementos: inicio, pasos, decisiones, bucles, fin, excepciones
3. Notación visual: diagramas de flujo agéntico
4. Patrones: secuencial, condicional, paralelo, iterativo, con escalado
5. Diseño de decisiones; manejo de errores: anticipar, recuperar, escalar

Unidad 6. El system prompt agéntico

1. El system prompt como código del agente
2. Diferencia prompt de usuario vs system prompt
3. Estructura de 7 secciones (identidad, contexto, comportamiento, herramientas, restricciones, incertidumbre, formato)
4. Técnicas: few-shot en system, cadenas de pensamiento; errores comunes; testear e iterar

Unidad Práctica 1. Especificar y diseñar un agente

Unidad 7. Guardrails: los 4 tipos

1. Filosofía: no confiar, verificar; diseñar para el peor caso
2. 4 tipos de guardrails: entrada (validar inputs, detectar manipulación), proceso (límites de pasos, timeouts, checkpoints), salida (validar output, confirmar irreversibles), recursos (tokens, rate limits, quotas)
3. Lista de comprobación pre-acción

Unidad 8. Seguridad de datos y contención

1. Protección de datos personales
2. Mitigación de prompt injection
3. Contención y respuesta ante fallos

Unidad 9. Escalado a humanos

1. Función del humano en el loop
2. Patrones de escalado operativo
3. Protocolos de intervención humana