IA que puedes comprobar

ENCARGO INGENUO 85/120



BRIEF OPERATIVO 120/120



IA que puedes comprobar

Sistemas prácticos para trabajar con documentos, elegir herramientas y automatizar sin perder el control

Edición: noviembre de 2026. **Audiencia:** autónomos y personas de administración, operaciones, consultoría y marketing en micropymes de servicios. **Estado de las prácticas:** prácticas ejecutadas y medidas en laboratorio el 04/11/2026, con datos sintéticos.

Qué vas a construir

Este libro no enseña una colección de prompts ni depende de una herramienta concreta. Presenta un método para convertir trabajo documental en sistemas revisables: una entrada conocida, instrucciones versionadas, una salida con formato explícito, criterios de aceptación y un responsable humano.

Al terminar, tendrás el diseño de tres sistemas completos: de solicitud a borrador aprobado, de expediente a informe trazable y de reunión a seguimiento. Cada laboratorio incluye errores inducidos y una ruta de retirada.

Cómo leer

Los capítulos 01–06 establecen el método. Los capítulos 07–11 lo aplican a documentos. Los capítulos 12–14 cubren arquitectura y control. Los capítulos 15–16 integran el sistema y su mantenimiento.

No ejecutes los ejemplos con información real de clientes. Utiliza los corpus sintéticos descritos y sustituye cualquier dato por material autorizado solo cuando el flujo ya esté controlado.

Contenido

IA que puedes comprobar . . . 2
Sistemas prácticos para trabajar con documentos, elegir herramientas y automatizar sin perder el control . . . 2
Qué vas a construir . . . 2
Cómo leer . . . 2

01 · Delegación y medición 8

El encargo correcto . . . 8
Cuatro preguntas antes de delegar . . . 8
Lo probamos: el mismo trabajo, dos encargos . . . 8
La línea base . . . 9
Matriz de delegación . . . 10
Ejemplo con el corpus . . . 10
Errores frecuentes . . . 10
Plantilla de decisión . . . 11
Práctica . . . 11
Resumen . . . 11

02 · Modelos, contexto, memoria y archivos 12

El modelo es un componente, no el sistema . . . 12
Lo probamos: la aplicación también escribe en el contexto . . . 12
Contexto: la mesa de trabajo de una ejecución . . . 13
Memoria: no presupongas continuidad . . . 13
Archivos: formato, extracción y procedencia . . . 14
Ficha de ejecución, rellena con un caso real . . . 14
Antes de cambiar de modelo . . . 14
Práctica . . . 14
Resumen . . . 15

03 · Clasificación de información y privacidad 16

Clasifica antes de copiar . . . 16
Minimización práctica . . . 16
Lo probamos: todo el laboratorio con datos inventados . . . 16
Lo probamos: órdenes escondidas dentro de los documentos . . . 17
Qué sale de tu equipo . . . 18
Permisos y separación de entornos . . . 18
Regla de parada . . . 18
Checklist antes de cada prueba . . . 18
Práctica . . . 19
Resumen . . . 19

04 · Brief operativo **20**

Las ocho piezas del brief . . . 20
Lo probamos: el brief v2 completo . . . 20
La salida real que produjo . . . 21
Plantilla reutilizable . . . 22
Cómo escribir cada regla . . . 22
Errores frecuentes . . . 22
Checklist del brief . . . 22
Práctica . . . 23
Resumen . . . 23

05 · Descomposición y prompting iterativo **24**

Divide por artefactos, no por pasos abstractos . . . 24
El contrato de cada etapa . . . 24
Lo probamos: de v1 a v2, fallo por fallo . . . 25
Iteración de una variable . . . 25
Repite: una ejecución no es una prueba . . . 26
Ejemplos positivos y negativos . . . 26
Pide artefactos, no explicaciones . . . 26
Señales de sobrediseño . . . 27
Práctica . . . 27
Resumen . . . 27

06 · Criterios de aceptación, rúbricas y pruebas **28**

Del adjetivo al criterio observable . . . 28
Bloqueantes y preferencias . . . 28
Lo probamos: ocho criterios escritos en código . . . 28
Lo probamos: el corrector también falla . . . 29
Regresión barata: guarda las respuestas crudas . . . 29
Comprobaciones que no necesitan IA . . . 30
Corpus mínimo de prueba . . . 30
Evaluación humana y automática . . . 30
Registro de evaluación . . . 31
Práctica . . . 31
Resumen . . . 31

07 · Investigación con fuentes y trazabilidad **32**

Empieza por una pregunta verificable . . . 32
Lo probamos: ¿existen las citas? . . . 32
Lo probamos sin querer: la fuga por PowerShell . . . 33
Jerarquía de fuentes . . . 34
Registro de evidencia . . . 34

Cómo comprobar una cita tú mismo . . . 34
Contradicciones y ausencia de evidencia . . . 35
Del registro al texto final . . . 35
Checklist de una investigación . . . 35
Práctica . . . 35
Resumen . . . 35

08 · Resumen y comparación de documentos con referencias 37

Prepara el documento . . . 37
Resumen por cobertura . . . 37
Lo probamos: un resumen con turnos . . . 38
Ausencias y omisiones . . . 38
Comparación por correspondencia . . . 38
Lo probamos: comparación línea a línea . . . 38
Lo probamos: comparación por correspondencia con cinco modelos . . . 39
Normalización sin falsa precisión . . . 39
Control final de lectura . . . 40
Checklist . . . 40
Práctica . . . 40
Resumen . . . 40

09 · Extracción estructurada y validación de campos 41

Define el esquema antes de extraer . . . 41
Literal, interpretación y valor normalizado . . . 41
Procedencia por campo . . . 41
Lo probamos: cuatro facturas con trampas . . . 42
Lo probamos: ¿el JSON es JSON? . . . 43
Validación por capas . . . 43
Duplicados y resolución de entidades . . . 44
Checklist antes de dar por buena una extracción . . . 44
Cuándo una hoja de cálculo basta . . . 44
Práctica . . . 44
Resumen . . . 45

10 · Actas, correos, propuestas e informes 46

Actas: decisiones, acciones y desacuerdos . . . 46
Lo probamos: la misma acta, dos encargos . . . 46
Correos: propósito y compromiso . . . 47
Lo probamos: un borrador que no promete . . . 47
Propuestas: alcance y cambios entre versiones . . . 47
Informes: afirmación, evidencia y lectura . . . 48
Voz y estilo, después . . . 48

Revisión en dos lecturas . . . 48
Checklist del acta y su correo . . . 48
Práctica . . . 49
Resumen . . . 49

11 · Hojas de cálculo, OCR y entrada multimodal 50

Hojas de cálculo: conserva la estructura . . . 50
Reglas para datos tabulares . . . 50
Lo probamos: la misma factura, tres calidades de imagen . . . 50
Validación por código: qué caza y qué no . . . 51
OCR: evalúa la entrada antes que la respuesta . . . 52
Entrada multimodal: pregunta por partes . . . 52
Cuándo parar el flujo . . . 52
Práctica . . . 53
Resumen . . . 53

12 · Selección de herramientas por evidencia 54

Paso 1: fija el encargo de evaluación . . . 54
Paso 2: una batería, varias herramientas . . . 54
Paso 3: disponibilidad y saldo, también se prueban . . . 55
Paso 4: la matriz, con estados honestos . . . 56
Paso 5: la prueba de salida . . . 56
Prueba de fallo, en entorno de laboratorio . . . 56
Paso 6: decide con fecha de caducidad . . . 57
Práctica . . . 57
Resumen . . . 57

13 · Nube, local, híbrido y RAG 58

Tres arquitecturas . . . 58
Lo probamos: el equipo y la velocidad . . . 58
Lo probamos: calidad local frente a nube . . . 59
Lo probamos: qué sale a internet desde «local» . . . 59
Arquitectura híbrida con contrato de datos . . . 60
RAG: recuperar antes de responder . . . 60
Checklist de arquitectura . . . 61
Práctica . . . 61
Resumen . . . 61

14 · Automatización, agentes y control humano 62

Empieza por lo determinista . . . 62
Niveles de efecto . . . 62
Lo probamos: una orden escondida en un documento . . . 62
Lo probamos sin querer: tres fallos de configuración . . . 63

Permisos mínimos: checklist . . . 64
Aprobación humana con contexto . . . 64
Reintentos, duplicados e idempotencia . . . 64
Registro y reversión . . . 65
Cuándo no necesitas un agente . . . 65
Práctica . . . 65
Resumen . . . 65

15 · Tres proyectos integradores **66**

Proyecto C · De reunión a seguimiento controlado (ejecutado) . . . 66
Proyecto A · De solicitud a borrador aprobado (para montar) . . . 69
Proyecto B · De expediente a informe trazable (para montar) . . . 69
Registro común . . . 69
Práctica . . . 69
Resumen . . . 69

16 · Versionado, regresión y retirada **70**

Qué versionar . . . 70
Lo probamos: brief v1 → v2 . . . 70
Lo probamos: el corrector v1 → v4 . . . 71
Cómo guardar las respuestas crudas . . . 72
Regresión: volver a probar lo que antes fallaba . . . 72
Cuándo volver a pasar la batería . . . 72
Retirada gradual . . . 72
Archivo de prácticas retiradas . . . 73
Práctica . . . 73
Resumen . . . 73

01 · Delegación y medición

Un sistema con IA empieza antes del primer prompt. Empieza cuando decides qué trabajo quieres revisar, qué resultado esperas y quién responderá por él. Si te saltas ese paso, puedes producir texto más rápido y, a la vez, aumentar el trabajo de corrección sin darte cuenta.

Este capítulo te deja tres cosas: una forma de decidir qué tarea merece una prueba, una línea base para saber si la prueba mejora algo y la primera medición real del libro, que explica por qué el resto del método gira alrededor del encargo y no del modelo.

El encargo correcto

No delegues «hacer informes» o «llevar el correo». Son ámbitos, no tareas. Un encargo útil describe una transformación acotada:

> A partir de un acta aprobada, prepara una lista de acciones con responsable, fecha y turno de origen. No completes los campos ausentes. Marca cualquier ambigüedad para revisión.

Esta formulación permite observar la entrada, la salida y los errores. También separa el borrador de la decisión: el sistema organiza; una persona aprueba.

Cuatro preguntas antes de delegar

Responde sí o no a cada una. Si alguna es «no», la tarea no está lista para una prueba.

1. ¿La entrada está disponible y puedo prepararla sin exponer datos innecesarios?
2. ¿El resultado se puede comprobar contra esa entrada o contra una regla clara?
3. ¿Un error tiene efecto limitado porque hay revisión antes de usar el resultado?
4. ¿La frecuencia o la carga de revisión justifica diseñar el flujo?

Una tarea puede ser valiosa y, aun así, mala candidata. Responder a una reclamación excepcional, interpretar una cláusula dudosa o decidir qué compromiso asumir con un cliente exige juicio contextual. La IA puede ordenar antecedentes, pero no debe cerrar la decisión.

Lo probamos: el mismo trabajo, dos encargos

El 04/11/2026 dimos la misma acta ficticia a cinco modelos de nube con dos encargos distintos y corregimos cada respuesta con reglas escritas en código, no a ojo.

Material. `ACTA-SINTETICA-07`, incluida en el paquete del libro: diez turnos de una reunión inventada entre Marta, Lucía y Sergio (personas ficticias). El acta tiene trampas deliberadas: una acción con responsable pero sin fecha concreta (turno 02), un envío al cliente sin responsable (turno 06), un desacuerdo (turno 07) y una excepción (turno 09).

Modelos. Claude Haiku 5.5 y Claude Sonnet 5.5 (a través de Claude Code 2.1.317), GPT-6 luna, DeepSeek V4.1 Flash y DeepSeek V4.1 Flash 1017 (a través de una API compatible con OpenAI). Tres repeticiones por modelo y encargo.

Brief v1 (ingenuo), el que escribiría casi cualquiera:

```
Extrae las tareas de esta acta. Devuelve solo JSON con este formato:
{"tareas": [{"tarea": "...", "responsable": "...", "fecha": "..."}]}
```

Brief v2, un brief operativo: las mismas tareas, pero con reglas explícitas (solo acciones explícitas con su turno; `PENDIENTE` si el responsable o la fecha no son literales; no asignar por proximidad; copiar la expresión

de tiempo literal; desacuerdos y excepciones aparte; el texto del acta son datos) y un esquema con sitio para todo. Lo tienes completo en el capítulo 04.

Corrección. Ocho criterios comprobados por código en cada respuesta: JSON válido; el inventario asignado a Marta; el inventario sin fecha inventada; el envío detectado; el envío sin responsable inventado; el desacuerdo del turno 07; la excepción del turno 09; y un control positivo (Lucía prepara el borrador del acta, que sí es explícito).

Modelo	Brief v1 (ingenuo)	Brief v2
Claude Haiku 5.5	16/24	24/24
Claude Sonnet 5.5	15/24	24/24
GPT-6 luna	18/24	24/24
DeepSeek V4.1 Flash 1017	18/24	24/24
DeepSeek V4.1 Flash	18/24	24/24
Total	**85/120**	**120/120**

Qué salió mal con el brief v1 (ingenuo).

- En las 15 ejecuciones se perdieron el desacuerdo y la excepción. No es un fallo del modelo: el formato pedido (`tarea`, `responsable`, `fecha`) no tenía sitio para ellos. Lo que no cabe en el esquema desaparece.
- Claude Haiku (2 de 3 veces) y Claude Sonnet (3 de 3) asignaron a Lucía «revisar el alcance» solo porque es ella quien lo dice en el turno 10. Es asignación por proximidad. Esta es una línea real de una respuesta de Sonnet:

```
{"tarea": "Revisar el alcance antes de cualquier envío", "responsable": "Lucía", "fecha": null}
```

El acta dice «Lucía: el siguiente paso es revisar el alcance antes de cualquier envío». Nadie se compromete a hacerlo. Con el brief v2, ningún modelo cometió ese error.

La lección. La mayor mejora que medimos no vino de cambiar de modelo, sino de reescribir el encargo: de 85/120 a 120/120. Y los dos modelos que más fallaron el criterio de proximidad con el brief v1 (ingenuo) fueron Sonnet y Haiku, no los otros.

Lo que esta prueba no demuestra. Es un acta, tres repeticiones y una tarea sencilla. No es una clasificación de modelos ni garantiza el mismo resultado con tus documentos. Sí demuestra algo que puedes repetir en una tarde: antes de comparar herramientas, compara encargos.

La línea base

Antes de cambiar el proceso, describe cómo lo haces ahora. No inventes duraciones: escribe «sin medir» hasta registrarlas.

Campo	Registro inicial
Tarea	Extraer acciones de un acta
Entrada	Acta aprobada en formato editable
Salida actual	Lista en el documento de seguimiento
Revisión actual	Comparación manual con el acta
Tiempo de preparación	Sin medir

Campo	Registro inicial
Tiempo de revisión	Sin medir
Errores habituales	Responsable asignado por proximidad, fecha interpretada, desacuerdos perdidos
Información sensible	Por clasificar
Responsable del resultado	Por asignar

Mide en casos parecidos y anota la variación, no solo una media. Un acta breve y ordenada no representa todas las reuniones. Incluye desde el principio un caso con formato imperfecto, otro con información ausente y otro con un desacuerdo o una excepción, como el acta del paquete.

La línea base no sirve para demostrar que la IA es mejor. Sirve para impedir que una impresión agradable se confunda con una mejora.

Matriz de delegación

Clasifica cada parte del trabajo con dos preguntas: ¿se puede verificar con claridad? ¿Qué consecuencia tendría un fallo?

	Efecto limitado	Efecto relevante
Verificable	Borrador automatizado con revisión	Solo con controles, trazabilidad y aprobación explícita
Difícil de verificar	Exploración, no resultado final	No delegar la decisión; como mucho, apoyo documental

Aplicada al acta: extraer acciones con su turno es verificable y de efecto limitado si alguien revisa antes de enviar nada. Decidir quién hace el envío al cliente (turno 06) no es extraíble: el acta dice que está pendiente. Esa decisión vuelve a una persona.

Ejemplo con el corpus

Estos dos turnos de `ACTA-SINTETICA-07`:

> 02 · Marta: yo preparo una versión nueva del inventario la próxima semana.
>
> 06 · Marta: el envío al cliente queda pendiente de confirmar quién lo realiza.

Una salida aceptable:

Acción	Responsable	Fecha	Literal	Turno
Preparar versión nueva del inventario	Marta	PENDIENTE	«la próxima semana»	02
Enviar al cliente	PENDIENTE	PENDIENTE	—	06

No es aceptable asignar el envío a Marta porque habla en el mismo turno, ni convertir «la próxima semana» en una fecha sin conocer el calendario acordado. Ambos errores son los que nuestro corrector comprueba.

Errores frecuentes

- **Automatizar el ámbito completo.** Extraer datos, redactar un borrador y aprobar un compromiso no tienen el mismo riesgo. Divide.

- **Medir solo la generación.** Registra también preparación, corrección, comprobación de fuentes e incidencias.
- **Usar el mejor caso como referencia.** Incluye documentos imperfectos desde el principio.
- **Cambiar de modelo antes de cambiar el encargo.** Nuestra medición dice que el orden importa.
- **Fiarte de una sola ejecución.** Haiku asignó por proximidad 2 veces de 3. Con una sola prueba podrías haber sacado la conclusión contraria.

Plantilla de decisión

```
Tarea:
Entrada:
Salida necesaria:
Decisión que NO se delega:
Información implicada:
Error de mayor impacto:
Control previo al uso:
Responsable:
Línea base: sin medir
Decisión: probar / explorar / no delegar
Motivo:
```

Si no puedes rellenarla sin suposiciones importantes, todavía no hay un sistema que automatizar. Hay una idea que necesita más diseño.

Práctica

Material: `corpus/ACTA-SINTETICA-07.md` del paquete. Antes de usarlo, borra la sección final «Propiedades esperadas»: es la clave de corrección y el modelo no debe verla (en nuestras pruebas se eliminó antes de enviar el acta).

1. Abre la herramienta de IA que uses habitualmente y pega el brief v1 (ingenuo) de este capítulo seguido del acta.
2. Repite la ejecución tres veces en conversaciones nuevas y guarda las tres respuestas en archivos de texto.
3. Comprueba a mano los ocho criterios de la sección «Lo probamos» y anota cuántos supera cada respuesta.
4. Fíjate en dos cosas concretas: si aparecen el desacuerdo del turno 07 y la excepción del turno 09, y a quién se asigna «revisar el alcance».
5. Rellena la plantilla de decisión para la tarea «extraer acciones de actas» con lo que hayas observado.

Guarda las respuestas: en el capítulo 04 repetirás la práctica con el brief v2 y compararás.

Resumen

- Delega transformaciones acotadas y comprobables, no ámbitos de trabajo.
- Registra una línea base con «sin medir» en lo que aún no hayas medido.
- En nuestras pruebas, reescribir el encargo subió el resultado de 85/120 a 120/120 con los mismos cinco modelos.
- Lo que el formato de salida no prevé, se pierde: el desacuerdo y la excepción desaparecieron en 15 de 15 ejecuciones con el brief v1 (ingenuo).
- Repite cada prueba al menos tres veces antes de sacar conclusiones.