Mecanismos de la IA generativa: caso Austral Café y Equipos

Actividad 07, Punto 2 · Curso Inmersión Agentes de IA para Negocios

Autor: Carlos Guillermo Paredes Pizarro | **Fecha:** 22 de octubre de 2026

Alcance, convenciones y resumen

El caso es una empresa **ficticia**, Austral Café y Equipos, que importa y comercializa café y equipos para cafeterías. Los cuatro documentos y los datos son **SIMULADOS**, creados para este ejercicio. Las IA recibieron los documentos pegados como texto en el chat: eso es un **contexto documental delimitado**, no un sistema RAG.

Cada resultado lleva su estado: **SIMULADO** (dato creado por nosotros), **EJECUTADO** (realmente corrido, con registro) o **PROPUESTO**. Cada respuesta se evaluó contra los documentos como **respaldada**, **no respaldada** (sin evidencia en los documentos, lo que no equivale a falsa) o **contradicha**. La coincidencia entre dos IA no se usó como prueba de acierto.

Prueba	IA y modelo mostrado (según lo informado)	Repeticiones	Estado
Consultas P1 a P6	ChatGPT “6.1 sol ligero”; Claude “Sonnet 5.5 medio”	1 por IA	EJECUTADA
Instrucciones de creatividad A, B y C	Las mismas dos IA	1 por variante y por IA	EJECUTADA
Tokenización	Herramienta web, opción “GPT-5.x & O1/3”	1	EJECUTADA
Temperatura 0.2, 0.7 y 1.5	Gemini 3.1 Pro Preview	1 por valor	EJECUTADA (valores informados)

Resumen de hallazgos. (1) Con documentos delimitados, ChatGPT y Claude respondieron con citas literales; no se detectó ninguna cita inventada. (2) Ambas se abstuvieron cuando faltaba el registro (P5) y advirtieron la contradicción documental (P6). (3) Se observaron pocos datos no respaldados, y ninguno basta para generalizar: una suposición de fecha (Claude, P5) y “nuestra bodega” (Claude, variante C). (4) Una discrepancia del registro que **no estaba prevista en el diseño** fue señalada por ChatGPT y Claude a partir de los documentos. Gemini la reprodujo porque el prompt la incluía entre los datos; no la detectó por sí mismo. (5) La creatividad pedida y la temperatura de Gemini cambiaron sobre todo la redacción; los hechos se conservaron en estas muestras. Todo con una ejecución por celda: describe lo observado y no permite generalizar.

1. Mapeo de áreas y documentos (SIMULADO)

Área	Documento	Tipo	Información que permite consultar
Logística	LOG-POL-01 Política de Despacho y Entregas (vigente desde 2026-09-01)	Política general	Plazos de despacho; recepción; plazo de reclamos por daños de transporte
Facturación y Finanzas	FIN-POL-02 Política de Facturación y Pagos (2026-05-15)	Política general	Emisión, plazo de pago, recargo por mora, imputación de pagos
Ventas y Servicio Técnico	VTS-POL-03 Condiciones de Venta, Garantía y Servicio Técnico (2026-02-01)	Política general	Cotizaciones, garantía, exclusiones, reclamo de daños en equipos
Logística, Facturación y Servicio Técnico	REG-04 Extracto del Registro de Gestión (extraído el 2026-10-15)	Registro de casos concretos	Estado de pedidos, facturas y solicitudes de clientes específicos

Las políticas dicen cómo debe ser; el registro dice qué ocurrió en un caso. Una política no confirma el estado de un pedido, un pago o una solicitud.

2. Consultas y fuente adecuada

N.º	Consulta (resumen)	Área	Fuente que permite responder	Qué pone a prueba
P1	Tiempo de despacho de 20 kg de café en grano tras confirmar el pago	Logística	LOG-POL-01, 1.1 (política); 1.3 para el tránsito	Respuesta explícita
P2	Factura vencida hace 6 días: ¿se cobra recargo?	Facturación y Finanzas	FIN-POL-02, 3.1 y 3.2 (política)	Respuesta explícita
P3	Molino de 8 meses al que le cayó agua: ¿lo cubre la garantía?	Ventas y Servicio Técnico	VTS-POL-03, 2.1, 2.2 y 3.1 (política)	Exclusión condicionada; dato ausente
P4	Pedido AC-2031: ¿cuándo pagar la factura y cuándo llega?	Logística y Facturación	REG-04 A-1 y B-3 (registro); FIN-POL-02, 2.1 (política)	Cruce de dos áreas
P5	Transferencia por la factura FA-0307: ¿ya está registrada?	Facturación y Finanzas	REG-04, Sección B: FA-0307 no figura; FIN-POL-02, 4.1 solo da un plazo general	Dato ausente
P6	Máquina recibida el 9 con carcasa golpeada, aviso el 13: ¿aceptan el reclamo?	Logística y Servicio Técnico	LOG-POL-01, 2.2; VTS-POL-03, 4.1; REG-04, A-2 y C-1	Contradicción documental

3. Tokenización y relaciones semánticas

3.1 Tokenización (EJECUTADA, prueba aparte)

Se usó una herramienta web de tokenización, independiente de los modelos comparados en las demás pruebas. Se registra tal como aparece en la captura (Anexo): opción seleccionada **“GPT-5.x & O1/3”**; no se atribuye ninguna codificación técnica, porque la captura no la muestra.

Frase	“El pedido de café está en tránsito.”
Resultado según la herramienta	8 tokens; 35 caracteres (verificado: la frase tiene 35 caracteres)
Segmentación (lista aportada por el usuario)	["El", " pedido", " de", " café", " está", " en", " tránsito", "."]
Lectura	En esta frase, con esta opción de esta herramienta, cada palabra común del español queda como una unidad con su espacio inicial, y el punto final como otra. Es una observación de una sola frase: no dice cómo se segmentan códigos, números ni otras palabras, ni cómo lo hace el tokenizador de otro modelo.

3.2 Relaciones semánticas (comparación cualitativa)

Valoración cualitativa propia: no se calcularon valores de similitud (no se usó ninguna herramienta de embeddings). Se distinguen tres ideas: **similitud semántica** (significados cercanos), **relación temática** (mismo tema o proceso, sin significar lo mismo) y **semejanza de escritura** (se parecen las letras, no el significado).

Par de términos	Relación	Lectura en el caso
despacho ↔ envío	Similitud semántica	LOG-POL-01 dice “despacho”; un cliente diría “envío”. Designan el mismo hecho: la salida del pedido.
reclamo ↔ queja	Similitud semántica parcial	En los documentos, “reclamo” es un trámite formal con plazo (48 horas o 5 días corridos); “queja” no lo implica. Tratarlos como equivalentes podría hacer pasar por alto el plazo.
factura ↔ recargo	Relación temática	El recargo se aplica sobre el saldo vencido de una factura (FIN-POL-02, 3.2), pero no significan lo mismo.
garantía ↔ reembolso	Relación temática (posventa)	VTS-POL-03 define la garantía (12 meses, defectos de fabricación); ningún documento regula reembolsos, así que una pregunta sobre reembolso no puede responderse con este material.
AC-2031 ↔ AC-2035	Semejanza de escritura	Los códigos difieren en un carácter y designan pedidos de clientes distintos (REG-04, A-1 y A-2). Confundirlos sería un error aunque se parezcan.

Par de términos	Relación	Lectura en el caso
recargo ↔ descuento	Opuestos en el mismo ámbito	Ambos ajustan un precio, pero con sentido contrario: se parecen por tema, no por significado. Que aparezcan juntos en un texto no los hace equivalentes.
pago confirmado ↔ sin pago registrado	Relación temática, con estados en conflicto	REG-04 usa ambas expresiones para el mismo pedido (A-1 y B-3). Son términos próximos que el registro presenta en conflicto: es la discrepancia que ChatGPT y Claude señalaron (INC-01).

4. Temperatura y creatividad

4.1 Experimento de temperatura numérica (EJECUTADO en Gemini, 1 repetición por valor)

Experimento **separado** de la comparación de estilos entre ChatGPT y Claude. Herramienta: Google AI Studio. Modelo: Gemini 3.1 Pro Preview, según lo informado. Las capturas disponibles muestran el modelo y el control de temperatura, pero no acreditan cada uno de los tres valores ejecutados. Cada valor se ejecutó en un chat nuevo, con el mismo prompt y las demás opciones iguales; solo cambió la temperatura. Los valores figuran **según lo informado**: no hay una captura de la configuración de cada ejecución, y las demás opciones no quedaron anotadas. El prompt, reproducido abajo, **entrega los datos ya extraídos** (incluida la discrepancia del registro) en lugar de los cuatro documentos, y fija tono, extensión y restricciones. Por eso este experimento mide variación de redacción con datos dados, no recuperación desde documentos.

```
Ejercicio académico. Empresa ficticia: Austral Café y Equipos.
Todos los datos siguientes son simulados.

DATOS DISPONIBLES
- Cliente: Cafetería Los Alerces.
- Pedido AC-2031: 40 kg de café en grano.
- Según el registro del 15/10/2026, fue despachado el 08/10/2026 y está en tránsito.
- Entrega estimada por el transportista: 16/10/2026. No existe confirmación de entrega.
- Factura FA-0311: emitida el 08/10/2026, vence el 29/10/2026.
- Hay 3 días corridos de gracia posteriores al vencimiento.
- Desde el cuarto día posterior al vencimiento se aplica un recargo único del 2% sobre el saldo vencido.
- El pedido figura con pago confirmado, pero la factura figura sin pago registrado. Esta discrepancia no está resuelta.

TAREA
Redacta un correo claro y cordial para informar al cliente sobre el pedido, la factura y la regla de recargo.

Usa únicamente los datos anteriores. No inventes información ni prometas gestiones. No solicites un nuevo pago como si la deuda estuviera confirmada. Distingue la entrega estimada de la confirmada y fecha el estado del pedido.

Entrega solo el correo, con asunto y firma, entre 90 y 120 palabras.
```

Temperatura	Palabras	Hechos (verificados contra los datos del prompt y los documentos)	Diferencias observadas
0.2	113	Fechas, código, 40 kg, 3 días de gracia y recargo de 2%: correctos	Saludo “Estimado equipo de…”; “Le recordamos que…”
0.7	113	Todos correctos	Saludo “Estimados Cafetería…” (concordancia dudosa); “Le informamos que…”
1.5	110	Todos correctos	Saludo “Estimada Cafetería…”; “Hay 3 días…”; separa algunas oraciones con punto y coma

Los tres correos conservan los mismos hechos y la misma estructura, y cumplen las restricciones del prompt: entre 90 y 120 palabras, solo el correo, estado fechado, entrega estimada distinguida de la confirmada y sin solicitar un nuevo pago. Los tres mencionan la discrepancia entre “pago confirmado” y “sin pago registrado”, que

el prompt incluía entre los datos. No se observaron datos falsos, promesas ni mayor creatividad. **Lectura prudente:** con una ejecución por valor no se puede concluir que una temperatura alta no genere errores ni que la baja los evite. Una explicación posible, no probada, es que un prompt tan acotado (datos entregados, tono, extensión y restricciones fijados) dejó poco margen de variación.

4.2 Comparación de instrucciones de creatividad (ChatGPT y Claude; no es temperatura)

COMPARACIÓN DE INSTRUCCIONES DE CREATIVIDAD. Se pidió el mismo correo sobre el pedido AC-2031 con tres instrucciones de estilo (A formal y literal, B clara y cordial, C creativa y original). No se atribuye ningún valor de temperatura a estas instrucciones.

Variante	ChatGPT	Claude
A	115 palabras; datos correctos; solo el correo	116 palabras; datos correctos; agrega notas no solicitadas
B	105 palabras; datos correctos; muy parecido a su variante A	113 palabras; datos correctos; notas no solicitadas; “Les recordamos” es una fórmula habitual de cortesía y no prueba una comunicación previa inventada
C	114 palabras; datos correctos; estilo más libre sin alterar hechos	118 palabras; hechos principales correctos, pero “ya salieron de nuestra bodega” NO está respaldado; notas no solicitadas

5. Comparación de respuestas y detección de alucinaciones

N.º	ChatGPT	Claude
P1	Respaldada	Respaldada
P2	Respaldada	Respaldada
P3	Respaldada. Condicional (“si el agua ingresó… está excluido”); no menciona los 12 meses	Respaldada, con reserva. Dice “probablemente no”, marcado como inferencia
P4	Respaldada. Señala la discrepancia como interna de REG-04	Respaldada. Señala la discrepancia y calcula bien el inicio del recargo (2026-11-02)
P5	Respaldada. Se abstiene	Se abstiene, pero afirma que el extracto es “anterior a la transferencia”: NO RESPALDADA
P6	Respaldada. Señala la contradicción; omite fechas y los 4 días	Respaldada. Da fechas y 4 días, y nota que una política es más reciente y la otra más específica

Lecturas. Las citas de ambas IA son literales. Ambas reconocieron lo que los documentos no permiten responder (P5) y advirtieron que 48 horas y 5 días corridos se contradicen sin decidir cuál prevalece (P6): la fecha más reciente (LOG-POL-01) y la mayor especificidad (VTS-POL-03) apuntan a reglas distintas. Las diferencias principales son de estilo: ChatGPT fue más breve y estricto con el formato; Claude fue más completo y agregó notas fuera del formato. Son observaciones de una ejecución y no establecen que una IA sea mejor.

6. Validación de resultados e inconsistencias

ID	Dónde	Hallazgo y evidencia	Clasificación
INC-01	REG-04, A-1 y B-3 (también A-2 y B-2)	REG-04 indica “Pago confirmado: 2026-10-06” para AC-2031 y “Pago: sin pago registrado” para su factura FA-0311. Señalada por ChatGPT y Claude en los documentos; Gemini la reprodujo porque el prompt la incluía.	Inconsistencia del material, no prevista en el diseño; no es alucinación
INC-02	LOG-POL-01, 2.2 y VTS-POL-03, 4.1	“48 horas desde la recepción” frente a “5 días corridos desde la recepción” para daños de transporte; ningún documento dice cuál prevalece.	Contradicción documental diseñada; ambas IA la señalaron
INC-03	Claude, P5	Afirma que el extracto es “anterior a la transferencia”; los documentos no dan la fecha actual ni la de la transferencia.	No respaldada

ID	Dónde	Hallazgo y evidencia	Clasificación
INC-04	Claude, creatividad C	"Ya salieron de nuestra bodega": REG-04 registra despacho y "En tránsito", pero no menciona ninguna bodega.	No respaldada
INC-05	ChatGPT y Claude, P3; ChatGPT, P4	"Hace falta un diagnóstico técnico" y "requiere conciliación" son necesidades de verificación razonables; los documentos no las establecen como procedimiento, y no se evalúan como protocolos de la empresa porque las respuestas no los afirman como tales.	Observación (no es un dato inventado)
INC-06	Claude, consultas y variantes A, B y C	Agrega notas fuera del formato; en notas cita "hoy (22 de octubre)", fecha ausente de los documentos.	Desviación de formato; dato externo en las notas
INC-07	Gemini, 0.2, 0.7 y 1.5	Los tres correos mencionan al cliente la discrepancia interna del registro. El prompt la incluía entre los datos y pedía no solicitar un nuevo pago como si la deuda estuviera confirmada, así que es coherente con el prompt; queda como decisión de comunicación a revisar.	Observación (no es un error)

Conclusiones

- Con documentos delimitados, las respuestas fueron comprobables: casi todo se pudo contrastar con un fragmento literal. Eso facilita la verificación, pero no garantiza una interpretación correcta.
- En estas muestras, las IA reconocieron lo ausente y lo contradictorio. Los pocos datos no respaldados fueron pequeños (una suposición de fecha y un adorno narrativo) y no permiten afirmar que una IA invente más que otra.
- La creatividad pedida y la temperatura de Gemini cambiaron sobre todo la redacción; los hechos se conservaron. No se concluye que una temperatura alta sea segura ni que una baja evite errores.
- El material propio tuvo una inconsistencia no prevista (INC-01). Revisar los documentos fuente es parte del control, y las IA ayudaron a detectarla.

Limitaciones

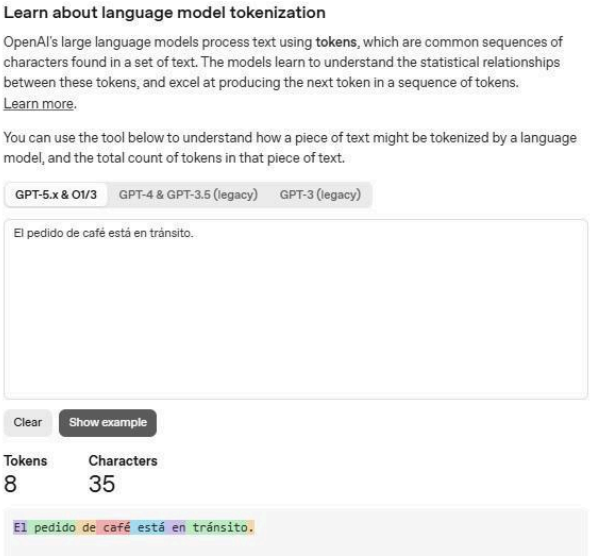
- **Una ejecución por celda** (IA, consulta, variante o temperatura): se describe lo observado; no hay generalización a otros modelos, tareas o empresas.
- **Contexto documental delimitado, no RAG.** La clave de evaluación es propia, y en P3 y P6 parte del juicio es interpretativo.
- **Controles de plataforma no verificables:** la desactivación de memoria y de búsqueda web no se pudo comprobar. Los modelos y las temperaturas figuran según lo informado, y no hay captura de la configuración de temperatura.
- **Modelos distintos en cada prueba** (ChatGPT, Claude, Gemini y la herramienta de tokenización): sus resultados no son comparables entre sí.
- **Tokenización:** una frase, una herramienta y una opción; la frase usada difiere de la propuesta en el material. **Semántica:** solo cualitativa, sin cálculo.
- **Gemini:** el prompt entrega los datos directamente (no los documentos) y fija tono, extensión y restricciones; por eso es otro experimento y su desempeño no se compara directamente con el de ChatGPT y Claude, que sí se comparan descriptivamente entre sí porque recibieron las mismas pruebas. Las otras opciones de la herramienta no quedaron registradas.

Reflexión. Una respuesta fluida y bien redactada no es, por eso, verdadera: la comprobación contra fuentes delimitadas y la lectura humana de lo que falta y de lo que se contradice siguen siendo necesarias, y aun así no eliminan los errores.

Anexo: evidencias

Archivos de respaldo (con las respuestas originales sin editar):

- **Registro_resultados_P1-P6.docx:** condiciones de cada plataforma; respuestas originales de ChatGPT y Claude (consultas y variantes A, B y C); tokenización; tres ejecuciones de Gemini; evaluación e inconsistencias.
- **Austral_Material_y_Clave.docx** y los archivos 01_bloque_comun_consultas.txt y 02A, 02B y 02C: documentos simulados, consultas, clave de evaluación (propuesta por nosotros, con una nota de revisión posterior a la ejecución) y bloques usados en los chats.
- **Captura de la herramienta de tokenización** (abajo).



Captura de pantalla de la herramienta de tokenización (sin editar).