Protocolo final de pruebas de estrés

Datos de esta lección

Fase	F6 — Proyecto completo
Módulo	M6.2 — Auditoría y graduación
Puntos de profesionalidad al superar el test	100

Objetivo práctico

Al terminar esta lección habrás sometido tu agente completo de Fontanería Ejemplo a una batería de quince categorías de pruebas de estrés, habrás rellenado una hoja de resultados con una puntuación objetiva por categoría, y sabrás decidir, con un criterio de aprobación claro, si el sistema está listo para entregarse a un cliente real o si necesita otra ronda de ajustes.

Resultado que obtendrás

Una hoja de resultados completa, con una puntuación total sobre 30 puntos, que documenta de forma objetiva el comportamiento de tu agente ante interrupciones, silencios, ruido, preguntas trampa, intentos de hacerlo inventar, urgencias, enfado, transferencias, cancelaciones, modificaciones, errores técnicos y caídas de proveedor.

Dónde encaja dentro del sistema final

Este «Protocolo final de pruebas de estrés» abre el módulo 6.2, la auditoría final del proyecto construido en la lección anterior: antes de considerar el sistema listo para un cliente real, se somete deliberadamente a las peores condiciones razonables, no sólo a una llamada de prueba tranquila.

Conocimientos previos

- Construcción completa del proyecto final (lección anterior), con el sistema de Fontanería Ejemplo funcionando de extremo a extremo.
- Casos límite: preparar al agente para lo inesperado (módulo 4.2).
- Transferencias de llamada y contingencias (módulo 4.2).
- Logs sin saturación y dónde están en cada plataforma (módulo 4.1).

Herramientas necesarias

- Tu proyecto completo de Fontanería Ejemplo, ya construido y funcional (lección anterior).
- Una hoja de cálculo o documento de texto para registrar la hoja de resultados.
- Un navegador con NVDA activo.
- Otra persona o un segundo teléfono, si es posible, para las pruebas de interrupción y de ruido simultáneo (opcional; también pueden simularse en solitario).

Posible coste

Cada prueba de esta lección es una llamada real: consume minutos de telefonía, tokens de LLM, transcripción y voz, según lo explicado en el módulo 2.1. Con quince categorías de prueba, el coste total de una ronda completa es notable comparado con una única llamada de prueba; revisa tu saldo antes de empezar y planea completar todas las categorías en una o dos sesiones, no llamada por llamada en días distintos, para poder comparar los resultados en las mismas condiciones.

Advertencia de seguridad

Ninguna prueba de esta lección debe implicar un riesgo real: la prueba de «olor a gas» o «inundación activa» consiste en decir esa frase durante la llamada de prueba, nunca en generar una situación de riesgo real. No provoques una caída real y prolongada de un proveedor de producción si tu cuenta está también en uso por otras pruebas del curso; usa una ventana de mantenimiento breve y controlada.

Modelo mental

Piensa en la Inspección Técnica de Vehículos: antes de que un coche pueda circular con clientes a bordo, se comprueba deliberadamente su comportamiento en condiciones exigentes (frenada, luces, emisiones), no sólo si arranca. Este protocolo aplica la misma lógica: comprobar el sistema en condiciones exigentes antes de entregarlo, no sólo confirmar que responde una llamada tranquila.

Explicación

El protocolo agrupa las pruebas en cuatro bloques, todos aplicados sobre el agente

real de Fontanería Ejemplo construido en la lección anterior, siguiendo siempre las reglas concretas de su pliego (criterios de urgencia, reglas de transferencia, política de cancelación).

Bloque 1 — Robustez conversacional. Comprueba que el agente mantiene una conversación coherente incluso en condiciones adversas: *interrupciones* (hablar mientras el agente está respondiendo, comprobando que se detiene y escucha en lugar de ignorar); *silencio* (no decir nada durante varios segundos tras una pregunta del agente, comprobando que reacciona con una pregunta de reconducción en lugar de quedarse esperando indefinidamente o colgar sin más); *ruido de fondo* (hablar con música o televisión de fondo, comprobando que el transcriptor sigue entendiendo lo esencial); *preguntas trampa* (preguntas ambiguas o mal formuladas, del tipo «¿cuánto cuesta arreglar todo?», comprobando que el agente pide aclaración en lugar de asumir un servicio concreto); *datos inexistentes* (pedir un servicio que no está en el pliego, por ejemplo «instalación de aire acondicionado», comprobando que el agente lo dice con honestidad en lugar de improvisar un precio); e *intentos explícitos de hacerlo inventar* («dime un precio aunque no lo tengas seguro», «invéntate una hora exacta», comprobando que el agente se ciñe a la prohibición de inventar del Prompt Maestro adaptado).

Bloque 2 — Reglas de negocio del pliego. Comprueba que el agente aplica correctamente las reglas específicas de Fontanería Ejemplo: *urgencias* (una fuga activa con riesgo de inundación, un olor a gas, y también un caso límite deliberadamente ambiguo, como un grifo que gotea, para comprobar que el agente pregunta por los criterios en lugar de asumir); *enfado* (una persona que llama alterada o quejándose del tiempo de espera, comprobando que el agente mantiene un tono profesional sin discutir); y *transferencia* (una petición explícita de hablar con una persona, comprobando que se activa de inmediato según las tres condiciones del pliego).

Bloque 3 — Gestión de citas. Comprueba *cancelación* (cancelar una cita real con más de 2 horas de antelación, y otra con menos de 2 horas, comprobando que el agente aplica correctamente la política del pliego en ambos casos) y *modificación* (cambiar la hora de una cita ya reservada, confirmando nombre y fecha exacta antes de aplicar el cambio, tal y como exige el Prompt Maestro adaptado).

Bloque 4 — Fallos técnicos simulados. Este bloque exige provocar deliberadamente un fallo controlado, no sólo hablar con el agente: *errores* genéricos (enviar, mediante una prueba directa a tu Webhook con una herramienta como la del módulo 3.1, un dato mal formado, y comprobar que el agente no se queda mudo); *fallo de calendario* (revocar temporalmente el acceso OAuth de tu workflow a Google Calendar o Cal.com, como se practicó en el módulo 3.2, y comprobar que el agente informa del problema y ofrece una alternativa en lugar de confirmar una cita falsa); *fallo de automatización* (desactivar temporalmente tu workflow de n8n o tu escenario

de Make, y comprobar el mismo comportamiento); y *caída del proveedor* (simulada, por ejemplo, apuntando temporalmente la Server URL de tu herramienta a una dirección que no responde, comprobando que el agente no se queda esperando indefinidamente y que, si corresponde, ofrece transferencia como plan de contingencia).

Sistema de puntuación. Cada una de las quince categorías (seis del Bloque 1, tres del Bloque 2, dos del Bloque 3, cuatro del Bloque 4) se puntúa de 0 a 2: **0** si el agente falla claramente (inventa un dato, ignora una regla del pliego, se queda sin respuesta, o confirma algo que no ha ocurrido realmente); **1** si el comportamiento es parcialmente correcto (por ejemplo, transfiere correctamente pero sin aplicar bien el fallback si la transferencia falla); **2** si el comportamiento es completamente correcto según el pliego. La puntuación máxima total es 30.

Criterio de aprobación. Siguiendo el mismo umbral del 80 % usado en el resto del curso para superar un test, este proyecto se considera listo para un cliente real con una puntuación mínima de 24 sobre 30. Por debajo de ese umbral, identifica las categorías con 0 o 1 punto, vuelve a la lección anterior para corregir la pieza correspondiente (prompt, herramienta, workflow o calendario, según el diagnóstico por descarte), y repite únicamente esas categorías, no la batería completa, antes de una nueva puntuación total.

Vocabulario nuevo

Prueba de estrés (agente conversacional)

Piénsalo así: La prueba de frenada de emergencia antes de sacar un coche a la carretera con pasajeros.

Prueba deliberada en condiciones adversas o poco frecuentes, diseñada para comprobar el comportamiento del sistema en sus límites, no en su uso más habitual.

Ejemplo: Provocar deliberadamente un silencio de diez segundos durante la llamada para comprobar cómo reacciona el agente.

Hoja de resultados

Piénsalo así: El acta de la inspección técnica, con cada punto revisado marcado como apto, apto con reservas o no apto.

Documento que registra, categoría por categoría, la puntuación obtenida y las observaciones concretas de cada prueba realizada.

Ejemplo: La fila «Urgencias» de la hoja de resultados anota una puntuación de 2 y la observación «El agente preguntó explícitamente por los tres criterios antes de clasificar la llamada».

Preparación

Ten preparado tu proyecto completo de Fontanería Ejemplo (lección anterior) y crea un documento nuevo o una hoja de cálculo con quince filas (una por categoría de prueba) y tres columnas: Categoría, Puntuación (0-2) y Observaciones.

Procedimiento paso a paso

1. **Contexto:** Vas a preparar tu hoja de resultados antes de empezar ninguna llamada de prueba.

   Acción de teclado: Crea un documento nuevo (por ejemplo, en Google Sheets, como en el módulo 5.4) y escribe en la primera fila las columnas: Categoría, Puntuación, Observaciones. En las quince filas siguientes, escribe el nombre de cada categoría de los cuatro bloques descritos en esta lección.

   Respuesta esperada de NVDA: NVDA leerá cada celda conforme la escribes.

   Qué significa: Tener la hoja preparada de antemano evita perder el hilo de qué prueba falta al pasar de una llamada a otra.

   Acción siguiente: Guarda el documento con un nombre identificable, por ejemplo «Pruebas de estrés — Fontanería Ejemplo».

2. **Contexto:** Vas a ejecutar el Bloque 1 (robustez conversacional): interrupciones, silencio, ruido, preguntas trampa, datos inexistentes e intentos de hacer inventar al agente.

   Acción de teclado: Realiza una llamada de prueba por cada una de las seis situaciones descritas en el Bloque 1 de esta lección, anotando en tu hoja la puntuación (0, 1 o 2) y una observación concreta para cada una.

   Respuesta esperada de NVDA: No aplica una única respuesta; escucha con atención cada reacción del agente durante la llamada.

   Qué significa: Este bloque comprueba la robustez del propio LLM y del prompt, sin implicar todavía a la automatización ni al calendario.

   Acción siguiente: No avances al Bloque 2 hasta haber puntuado las seis

categorías del Bloque 1.

3. **Contexto:** Vas a ejecutar el Bloque 2 (reglas de negocio): urgencias, enfado y transferencia.

   Acción de teclado: Realiza una llamada de prueba para una urgencia real, otra para un caso límite ambiguo (por ejemplo, un grifo que gotea), otra simulando enfado, y otra pidiendo explícitamente hablar con una persona, anotando la puntuación y la observación de cada una.

   Respuesta esperada de NVDA: No aplica una única respuesta; escucha si el agente aplica correctamente los criterios exactos del pliego.

   Qué significa: Este bloque comprueba si el Prompt Maestro adaptado refleja con precisión las reglas de negocio reales del cliente.

   Acción siguiente: Anota especialmente si el agente pregunta explícitamente por los criterios de urgencia en el caso ambiguo, en lugar de asumirlos.

4. **Contexto:** Vas a ejecutar el Bloque 3 (gestión de citas): cancelación y modificación.

   Acción de teclado: Reserva primero una cita de prueba con más de 2 horas de antelación. Después, realiza una llamada cancelándola, otra reservando una segunda cita para dentro de menos de 2 horas e intentando cancelarla, y una tercera modificando la hora de una cita ya reservada.

   Respuesta esperada de NVDA: No aplica una única respuesta; confirma en tu calendario real si cada cambio se ha aplicado correctamente tras cada llamada.

   Qué significa: Este bloque comprueba tanto la herramienta cancelar_cita como la aplicación correcta de la política de las 2 horas del pliego.

   Acción siguiente: Anota la puntuación y la observación de ambas categorías antes de continuar.

5. **Contexto:** Vas a ejecutar el Bloque 4 (fallos técnicos simulados), provocando deliberadamente cada fallo antes de llamar.

   Acción de teclado: Para «fallo de calendario», revoca temporalmente el acceso OAuth de tu workflow y realiza una llamada de prueba pidiendo una cita; para «fallo de automatización», desactiva temporalmente tu workflow o escenario y repite la prueba; para «caída del proveedor», cambia temporalmente la Server URL de una herramienta a una dirección inválida y repite la prueba. Restaura

cada configuración inmediatamente después de cada prueba.

Respuesta esperada de NVDA: No aplica una única respuesta; escucha si el agente informa del problema con honestidad y ofrece una alternativa razonable en lugar de confirmar algo que no ha ocurrido.

Qué significa: Este bloque comprueba el comportamiento del sistema en su peor caso razonable, el más exigente de los cuatro bloques.

Acción siguiente: Restaura todas las configuraciones a su estado normal antes de calcular la puntuación total.

6. **Contexto:** Vas a calcular la puntuación total y aplicar el criterio de aprobación.

   Acción de teclado: Suma las quince puntuaciones de tu hoja de resultados (usando una fórmula de suma en tu hoja de cálculo, si la usas) para obtener el total sobre 30.

   Respuesta esperada de NVDA: NVDA leerá el resultado de la fórmula al recorrer la celda correspondiente.

   Qué significa: Una puntuación de 24 o más sobre 30 (80 %) indica que el proyecto está listo para un cliente real; por debajo, identifica las categorías con 0 o 1 punto.

   Acción siguiente: Si hay categorías por debajo del umbral, vuelve a la lección anterior para corregir la pieza concreta señalada por el diagnóstico por descarte, y repite sólo esas categorías antes de una nueva puntuación total.

Errores frecuentes y diagnóstico

El agente responde bien a todas las pruebas conversacionales del Bloque 1, pero falla en el Bloque 4 (fallos técnicos simulados).

Cómo localizarlo: Es un patrón habitual: un prompt bien escrito resuelve la robustez conversacional, pero no protege por sí solo frente a un fallo real de la automatización o el calendario si no existe un manejo de errores adecuado en el workflow.

Solución: Revisa que tu workflow de n8n o escenario de Make devuelva siempre una respuesta (de éxito o de error explícito) en un tiempo razonable, en lugar de dejar la petición sin respuesta cuando algo falla.

Al simular un fallo de calendario, el agente confirma en voz alta una cita que en realidad no se ha creado.

Cómo localizarlo: El agente está confirmando antes de recibir una respuesta de éxito real de la herramienta, en lugar de esperar el resultado.

Solución: Revisa la instrucción del Prompt Maestro adaptado sobre citas y confirmaciones: el paso de confirmación debe ocurrir después de recibir la respuesta de la herramienta, nunca antes.

La puntuación total queda justo por debajo del umbral de aprobación (por ejemplo, 22 sobre 30).

Cómo localizarlo: No es necesariamente un fallo grave: puede deberse a una o dos categorías puntuales con 0 o 1 punto, no a un fallo generalizado del sistema.

Solución: Identifica exactamente qué categorías puntuaron bajo, corrige sólo la pieza correspondiente según el diagnóstico por descarte, y repite únicamente esas categorías antes de recalcular el total.

Comprobación final

Completa la hoja de resultados de las quince categorías con una puntuación y una observación para cada una, calcula el total sobre 30, y confirma si alcanzas el umbral de aprobación de 24 puntos. Si lo alcanzas, la comprobación es positiva y el proyecto está listo para el plan de 30 días de la siguiente lección; si no lo alcanzas, la comprobación es igualmente válida en tanto que diagnóstico honesto, y el paso siguiente es corregir las categorías señaladas.

Qué acabas de conseguir

Tienes una auditoría objetiva y documentada del comportamiento de tu sistema completo en sus condiciones más exigentes, con una puntuación y un criterio de aprobación claros, en lugar de una impresión subjetiva de que «parece que funciona bien».

Ejercicio práctico

Elige la categoría con la puntuación más baja de tu hoja de resultados y describe, por escrito, qué pieza concreta del sistema (prompt, herramienta, workflow o calendario) corregirías primero según el diagnóstico por descarte, antes de repetir esa prueba.

Resumen de lo imprescindible

- El protocolo agrupa quince categorías de prueba en cuatro bloques: robustez conversacional, reglas de negocio del pliego, gestión de citas y fallos técnicos simulados.
- Cada categoría se puntúa de 0 a 2, con una puntuación máxima total de 30.
- El umbral de aprobación es 24 sobre 30 (80 %), coherente con el umbral usado en el resto del curso para superar un test.
- Por debajo del umbral, se corrige únicamente la pieza señalada por el diagnóstico por descarte y se repiten sólo las categorías afectadas, no la batería completa.
- El Bloque 4 exige provocar fallos técnicos de forma deliberada y controlada, no sólo probar el sistema en condiciones ideales.

No necesitas aprender todavía

Todavía no necesitas diseñar tu estrategia comercial para conseguir el primer cliente: es el contenido de la siguiente lección.

Estoy preparado para continuar si puedo…

- Ejecutar una batería completa de pruebas de estrés sobre un agente telefónico ya construido.
- Puntuar de forma objetiva el comportamiento del sistema con una rúbrica de 0 a 2 por categoría.
- Aplicar un criterio de aprobación claro (24 sobre 30) para decidir si un proyecto está listo para un cliente real.
- Provocar fallos técnicos controlados (calendario, automatización, proveedor) para comprobar el comportamiento del sistema en su peor caso razonable.

Test de comprobación

> *Para registrar el resultado y obtener Puntos de profesionalidad, realiza este test desde la versión web de la lección. Las opciones se muestran aquí sin señalar la respuesta correcta.*

1. Tu agente responde perfectamente a las seis pruebas del Bloque 1 (robustez conversacional) e incluso gestiona bien el enfado y las transferencias del Bloque 2, pero al simular una caída del proveedor de automatización, la llamada se queda en silencio indefinidamente sin que el agente diga nada. Según el criterio de puntuación de esta lección, ¿qué categoría concreta debe reflejar esa

puntuación baja?

   a. La categoría «caída del proveedor» del Bloque 4, con una puntuación de 0, sin que eso afecte necesariamente a la puntuación ya obtenida en los Bloques 1 y 2.
   b. Todas las categorías de los cuatro bloques deben puntuarse de nuevo con 0, porque un solo fallo invalida toda la ronda de pruebas anterior.
   c. La categoría «silencio» del Bloque 1, porque cualquier silencio durante una llamada pertenece siempre a esa categoría.

2. Tu hoja de resultados suma 21 puntos sobre 30. Según el criterio de aprobación de esta lección, ¿cuál es la actuación correcta?

   a. Identificar las categorías con puntuación baja, corregir únicamente la pieza señalada por el diagnóstico por descarte en cada una, y repetir sólo esas categorías antes de una nueva puntuación total.
   b. Dar el proyecto por terminado igualmente, porque 21 puntos es una puntuación razonablemente alta.
   c. Reconstruir el proyecto completo desde cero, sin usar ningún dato de la hoja de resultados.

3. La prueba de urgencias incluye deliberadamente un caso ambiguo (un grifo que gotea sin riesgo de inundación) en lugar de sólo casos claros (una fuga activa o un olor a gas). ¿Por qué esta lección considera necesario incluir ese caso ambiguo?

   a. Porque comprueba si el agente pregunta explícitamente por los criterios de urgencia del pliego antes de clasificar la llamada, en lugar de asumir la urgencia por el tono de voz o por la palabra «grifo».
   b. Porque un grifo que gotea es, según el pliego, siempre una urgencia real que exige transferencia inmediata.
   c. Porque las pruebas ambiguas no tienen ningún valor diagnóstico y sólo sirven para alargar la batería de pruebas.

4. Durante la prueba de fallo de calendario, el agente dice en voz alta «tu cita ha quedado confirmada» aunque el evento nunca llegó a crearse por el fallo simulado. ¿Qué corrección concreta señala esta lección para ese fallo?

   a. Revisar que el Prompt Maestro adaptado sólo confirme una cita después de recibir la respuesta real de éxito de la herramienta, nunca antes.
   b. Eliminar por completo la herramienta reservar_cita del proyecto, ya que ha demostrado no ser fiable.
   c. Cambiar de proveedor de voz, porque una confirmación incorrecta siempre indica un problema de texto a voz.

Fuentes

- Automatically test your agent — Retell AI (consultado el 2026-08-30)

Última actualización de este contenido: **2026-08-30**.