

PLIEGO DE REQUERIMIENTOS Y ESPECIFICACIONES TÉCNICAS

DESAFÍO DE IMPLEMENTACIÓN DE AGENTES DE IA

Participantes:

- **Néstor Raúl Sáenz Chajín**: Implementación basada en LangGraph.
- **Leonardo David Scelza**: Implementación basada en synapseForge.

---

1. OBJETO DEL DESAFÍO

El presente pliego establece los requerimientos, condiciones técnicas, restricciones, metodología de evaluación y criterios de adjudicación correspondientes al Desafío de Implementación de Agentes de IA.

El objetivo consiste en determinar cuál de las dos implementaciones es capaz de construir y ejecutar el agente más eficiente, preciso, consistente y adecuado para resolver consultas basadas en una misma Base de Conocimiento, bajo condiciones técnicas equivalentes.

Ambos participantes deberán resolver exactamente los mismos casos de prueba, utilizando los mismos modelos de lenguaje y de embeddings, y respetando las restricciones establecidas en este documento. El desafío evalúa la calidad de la implementación y la eficiencia en el uso de los recursos, no la superioridad inherente de un framework sobre otro.

2. CRONOGRAMA

- **Fecha de inicio:** 01/10/2026
- **Fecha de finalización:** 20/10/2026

Las fechas indicadas podrán ser modificadas únicamente de común acuerdo entre los participantes y la organización.

3. PREMIO Y RECONOCIMIENTO

El participante que resulte ganador será acreedor de:

- **Certificado Oficial de Ganador** emitido por **synapse.ai (SYNAPSE AI SAS)**, acreditando la excelencia técnica y el primer puesto en el Desafío de Implementación de Agentes de IA.
- Reconocimiento público en los canales oficiales de la organización.

4. BASE DE CONOCIMIENTO

Ambas implementaciones deberán trabajar sobre la misma fuente de información denominada: **Flores del Valle — Base de Conocimiento**

La Base de Conocimiento contiene información relativa a la empresa ficticia Flores del Valle S.A.S., incluyendo: información general, sucursales, horarios, reservas, cancelaciones, reprogramaciones, pagos, entregas, cumpleaños, matrimonios, funerales, otros eventos, descuentos por volumen, preguntas frecuentes, y restricciones y excepciones comerciales.

Esta información constituye la fuente de verdad del desafío. Los agentes deberán fundamentar sus respuestas exclusivamente en dicha información. La Base de Conocimiento será entregada mediante un enlace independiente al presente documento.

5. CASOS DE PRUEBA

Se establecen tres consultas de dificultad progresiva. Cada participante deberá ejecutar exactamente las mismas consultas sobre su implementación.

5.1. CASO 1 — DIFICULTAD BAJA

- **Consulta:** ¿Cuál es el horario de atención de la Sede Norte de Flores del Valle y qué atención ofrece los domingos?
- **Objetivo:** Evaluar recuperación básica de información, comprensión de horarios, identificación de excepciones y precisión de la respuesta.

5.2. CASO 2 — DIFICULTAD MEDIA



- **Consulta:** Quiero organizar un cumpleaños para 60 personas. ¿Qué opciones de cumpleaños tienen para esa cantidad de invitados, cuánto costaría el servicio y qué descuento corresponde aplicar?
- **Objetivo:** Evaluar recuperación de información comercial, interpretación de cantidades, aplicación de reglas de descuento, interpretación del umbral de volumen, cálculo y combinación de información proveniente de distintas secciones de la Base de Conocimiento.

El agente deberá interpretar correctamente que: "más de 50" significa 51 o más; el descuento correspondiente para cumpleaños es del 10%; el descuento se aplica sobre el valor del pedido (no sobre el costo de entrega); y los descuentos por volumen no son acumulables con promociones.

5.3. CASO 3 — DIFICULTAD ALTA

- **Consulta:** Quiero organizar un matrimonio para 250 invitados dentro de 20 días. Me interesa el paquete Elegance, necesito agregar 5 centros de mesa adicionales y quiero saber cuánto tendría que pagar, qué descuento me corresponde, cuánto debo entregar como anticipo, si puedo pagar en cuotas y si puedo solicitar una entrega urgente. Indicame también qué condiciones y restricciones debería tener en cuenta.
- **Objetivo:** Evaluar comprensión de múltiples reglas, detección de incompatibilidades, aplicación de restricciones, interpretación de umbrales, cálculo, descuentos, anticipos, modalidades de pago, condiciones de reserva y entrega.

El agente deberá detectar, entre otras cuestiones, que: las reservas para matrimonios requieren un mínimo de 30 días (por lo tanto, la solicitud de 20 días es incompatible con la política); para matrimonios de más de 200 invitados corresponde un 20% de descuento; el paquete Elegance tiene un valor de $2.500.000; cada centro de mesa adicional tiene un valor de $35.000; los matrimonios pueden abonarse hasta en 3 cuotas, sujeto a aprobación; el anticipo requerido es del 50%; y una entrega urgente implica un incremento del 20% del valor del pedido, está sujeta a disponibilidad y corresponde únicamente cuando se solicita con menos de 4 horas de anticipación. *La respuesta deberá diferenciar claramente entre las condiciones que pueden calcularse y las condiciones que invalidan la solicitud original.*

5.4. CASOS 4 A 11 — CONSULTAS ADICIONALES DEL JURADO

Se establecen ocho consultas adicionales de autoría del jurado humano. Cada uno de los 4 integrantes aportará 2 consultas, las que deberán ser resueltas por ambos participantes sobre su implementación.

- **Consulta:** A definir por cada integrante del jurado humano al momento de la evaluación.
- **Objetivo:** A indicar por cada integrante del jurado humano al momento de proponer cada consulta, precisando qué se busca evaluar con la misma.

Dichas consultas no serán conocidas previamente por los participantes y serán develadas en el momento de la evaluación.

5.5. CASOS 12 A 15 — CONSULTAS CRUZADAS ENTRE PARTICIPANTES

Se establecen cuatro consultas cruzadas entre participantes. Cada participante propondrá 2 consultas para que sean resueltas por su adversario sobre su implementación.

- **Consulta:** A definir por cada participante al momento de la evaluación.
- **Objetivo:** A indicar por cada participante al momento de proponer cada consulta, precisando qué se busca evaluar con la misma.

Dichas consultas no serán conocidas previamente por el adversario ni por el jurado y serán develadas en el momento de la evaluación.

6. MODELOS DE LENGUAJE Y EMBEDDINGS

Ambos participantes deberán utilizar exactamente los mismos modelos, bajo las mismas condiciones de acceso (Free Tier):

- **Modelo de Lenguaje (LLM):** Gemini 3.5 Flash Lite
- **Modelo de Embeddings:** Gemini Embedding 2

No se permitirá que un participante utilice un modelo superior, diferente o adicional para resolver los casos de prueba. El modelo utilizado deberá permanecer constante durante todo el desafío.

7. PARÁMETROS DEL MODELO

Los parámetros de generación del modelo (temperatura, `top_p`, `top_k`, límite de tokens de salida, etc.) serán de libre elección para cada participante. La elección de parámetros forma parte de la estrategia de cada implementación y deberá quedar registrada al momento de la evaluación.



8. ENTORNO DE EJECUCIÓN Y ENTREGA

- Ambos agentes deberán ejecutarse localmente.
- La solución deberá ser accesible mediante un navegador web local.
- **Entrega y Trackeo:** Todas las respuestas generadas, junto con el trackeo completo de la ejecución del agente (logs de pasos, llamadas a herramientas, tiempos, etc.), deberán ser almacenados en una base de datos **SQLite (formato `.db`)** para garantizar la compatibilidad con la plataforma de evaluación.
- Dicho archivo `.db` y los resultados deberán subirse a **challenges.synapseaihub.com.ar**.

9. RESTRICCIÓN DE EJECUCIÓN (SIN PARALELIZACIÓN)

Queda expresamente prohibida la paralelización de procesos o llamadas al modelo durante la resolución de los casos de prueba. Todas las operaciones deberán ejecutarse de manera secuencial para garantizar una condición comparable y evitar ventajas mediante concurrencia.

- **Permitido:** Ejecución secuencial de subagentes, llamadas secuenciales al LLM, herramientas, RAG, memoria, recuperación de información, procesamiento local, cálculos y validaciones.
- **No permitido:** Llamadas simultáneas al LLM, ejecución paralela de subagentes, múltiples ramas ejecutándose concurrentemente o cualquier mecanismo cuyo objetivo sea reducir el tiempo total mediante concurrencia.

10. ORQUESTACIÓN

La estrategia de orquestación queda a criterio de cada participante. Se permitirá utilizar agentes, subagentes, herramientas, RAG, memoria, recuperación de contexto, planificación, razonamiento, validación, procesamiento estructurado, workflows o cualquier mecanismo propio del framework utilizado, siempre que se respeten las restricciones generales.

11. RESTRICCIÓN SOBRE INFORMACIÓN EXTERNA

La Base de Conocimiento constituye la fuente oficial. No se permitirá utilizar información externa de Internet (buscadores, páginas web externas, APIs externas de conocimiento) para completar, corregir o sustituir información contenida en la Base de Conocimiento. El acceso al modelo de lenguaje y embeddings establecido constituye la única excepción permitida.

12. MÉTRICAS Y SISTEMA DE EVALUACIÓN

Cada dimensión se evaluará bajo un sistema ternario determinista, eliminando la subjetividad. Los estados posibles y su valor asociado son:

- **CUMPLE:** Asigna el 100% del peso de la dimensión.
- **PARCIAL:** Asigna el 50% del peso de la dimensión.
- **NO CUMPLE:** Asigna el 0% del peso de la dimensión.

Fórmula de cálculo: `Puntuación Final = Sumatoria (Estado x Peso)`

Dimensión	Peso	CUMPLE (100%)	PARCIAL (50%)	NO CUMPLE (0%)
Calidad de respuesta	25%	Responde el 100% de los puntos solicitados, cero alucinaciones y cita la fuente en la Base de Conocimiento.	Responde más del 80% de los puntos y no presenta alucinaciones.	Responde menos del 80% de los puntos o presenta al menos una alucinación.
Razonamiento y reglas	18%	Detecta todas las incompatibilidades y los cálculos son 100% correctos.	Detecta la incompatibilidad principal pero falla en un cálculo menor.	Ignora la incompatibilidad principal o presenta un error de cálculo mayor al 5%.
Eficiencia de orquestación	15%	Caso 1: 2 llamadas o menos. Caso 2: 3 llamadas o menos. Caso 3: 4 llamadas o menos.	Caso 1: 3 llamadas. Caso 2: 4 o 5 llamadas. Caso 3: 5 o 6 llamadas.	Caso 1: 4 llamadas o más. Caso 2: 6 llamadas o más. Caso 3: 7 llamadas o más.
Eficiencia de recursos	15%	Caso 1: 1.000 tokens o menos. Caso 2: 2.500 tokens o menos. Caso 3:	Caso 1: entre 1.001 y 2.000 tokens. Caso 2: entre 2.501 y	Caso 1: más de 2.000 tokens. Caso 2: más de 4.000 tokens.



		4.000 tokens o menos.	4.000 tokens. Caso 3: entre 4.001 y 6.000 tokens.	Caso 3: más de 6.000 tokens.
Latencia LLM	12%	Caso 1: 1,8 segundos o menos. Caso 2: 3,6 segundos o menos. Caso 3: 6,0 segundos o menos.	Caso 1: entre 1,9 y 3,6 segundos. Caso 2: entre 3,7 y 7,2 segundos. Caso 3: entre 6,1 y 12,0 segundos.	Caso 1: más de 3,6 segundos. Caso 2: más de 7,2 segundos. Caso 3: más de 12,0 segundos.
Consistencia	10%	5 de 5 ejecuciones idénticas en respuesta y reglas aplicadas.	4 de 5 ejecuciones correctas con variación menor en formato.	3 de 5 ejecuciones correctas o menos, o varianza de tokens mayor al 15%.
Cumplimiento de restricciones	5%	Cero violaciones, archivo .db completo y modelos correctos.	1 violación no crítica (ejemplo: campo mal nombrado en el .db).	1 o más violaciones críticas (uso de otro modelo, web search, paralelización).

Nota técnica: Los rangos de latencia incluyen un coeficiente de seguridad de 1,2x sobre los valores teóricos del modelo para absorber la variabilidad de red y carga de servidores.

13. REPETICIÓN DE LAS PRUEBAS

Cada caso podrá ejecutarse múltiples veces para reducir el impacto de la variabilidad propia de los modelos de lenguaje. La cantidad definitiva de ejecuciones será establecida antes de comenzar la evaluación y será idéntica para ambos participantes.

14. JURADO

El resultado será determinado por un jurado compuesto por:

- **Jurado humano:** 4 integrantes (2 designados por Néstor Raúl Sáenz Chajín y 2 designados por Leonardo David Scelza).
- **Jurado artificial:** 1 LLM independiente como juez, que utilizará una rúbrica previamente definida y recibirá las mismas condiciones de evaluación para ambas implementaciones.

15. EVALUACIÓN CIEGA

Siempre que resulte técnicamente posible, las implementaciones serán presentadas al jurado como "Agente A" y "Agente B", sin revelar inicialmente cuál corresponde a LangGraph y cuál a synapseForge, con el objetivo de reducir sesgos. La identificación podrá revelarse una vez finalizada la evaluación.

16. IGUALDAD DE CONDICIONES

Ambos participantes dispondrán de la misma Base de Conocimiento, los mismos modelos, el mismo nivel de acceso, la misma cantidad de casos, consultas, restricciones y condiciones de evaluación.

17. LIBERTAD DE IMPLEMENTACIÓN

Fuera de las restricciones expresamente establecidas, cada participante podrá utilizar libremente las capacidades de su framework (LangGraph o synapseForge). Se evaluará la solución integral desarrollada, no solo la capacidad de realizar llamadas al LLM.

18. CRITERIO GENERAL DE VICTORIA

El ganador será el participante cuya implementación obtenga la mayor Puntuación Final, calculada mediante la sumatoria de los pesos asignados por estado en cada dimensión de evaluación. La prioridad será producir respuestas correctas y completas, manteniendo simultáneamente la mayor eficiencia posible.

19. PRINCIPIO DE EVALUACIÓN

El desafío evaluará la capacidad de cada implementación para resolver problemas reales de un agente basado en conocimiento. Se privilegiará la capacidad de: interpretar la consulta, recuperar información relevante, aplicar reglas, detectar excepciones, razonar sobre restricciones, calcular cuando corresponda y construir una respuesta correcta.



20. CRITERIO ANTE INFORMACIÓN CONTRADICTORIA O INSUFICIENTE

Cuando una consulta contenga condiciones incompatibles con la Base de Conocimiento, el agente deberá identificar dicha incompatibilidad. No se considerará correcta una respuesta que ignore una restricción para satisfacer literalmente la solicitud. El agente deberá indicar que la solicitud no es válida bajo las condiciones actuales, explicar qué regla lo impide y, si corresponde, realizar cálculos hipotéticos claramente identificados como tales.

21. REGISTRO DE RESULTADOS

Para cada ejecución deberán registrarse en el archivo `.db` de trackeo, como mínimo: Caso, Agente, Respuesta, Cantidad de llamadas LLM, Tokens de entrada, Tokens de salida, Tokens totales, Latencia total LLM, Costo normalizado, Cumplimiento de restricciones, Evaluación humana, Evaluación LLM-as-Judge, Calificación obtenida en cada dimensión y Puntuación ponderada final.

22. TRANSPARENCIA

Cada participante deberá estar en condiciones de demostrar, ante el jurado, cómo su implementación obtuvo el resultado presentado, mostrando el flujo de ejecución, llamadas al modelo, herramientas utilizadas, recuperación de información, cantidad de llamadas, consumo de tokens, tiempos registrados y resultado final, a través del archivo de trackeo proporcionado.

23. RESTRICCIONES GENERALES

Queda expresamente prohibido: Utilizar un modelo diferente o adicional al establecido; consultar fuentes externas de información; utilizar paralelización; alterar la Base de Conocimiento; modificar las consultas proporcionadas; manipular deliberadamente las métricas; ocultar llamadas al modelo; o diseñar la solución para responder mediante valores precargados o hardcodeados.

24. REGLA DE INTERPRETACIÓN

Ante cualquier duda de interpretación del presente pliego, deberá priorizarse el siguiente orden: igualdad de condiciones entre participantes, cumplimiento de las restricciones, reproducibilidad de la evaluación, objetividad de las métricas y calidad de la solución.