BRIDGEWARE
Technologies

Ficha de Modelo IA

Proyecto: Agente Inteligente de WhatsApp GTN v.1.0
Junio 2025

Desarrollador:
Bridgeware Technologies

Esta ficha del modelo describe el marco operativo y ético del Agente Inteligente de WhatsApp de GTN v.1.0. Este agente actúa como un asistente de IA de primera línea para consultas sobre alquiler de vehículos y turismo, cerrando la brecha entre la recuperación de información automatizada y el soporte humano personalizado.

Detalles del Modelo

Arquitectura Principal

- Plataforma: Plataforma de automatización de flujos de trabajo n8n.
- Motor LLM: GPT-4.1-mini a través de la API de OpenAI.
- Comunicación: API de WhatsApp Business de Meta.
- Integraciones: Gmail a través de la API de Google Cloud (para escalación de clientes potenciales).

Tipo de Entrenamiento

- Marco de trabajo (Framework): Model Context Protocol (MCP) utilizado para exponer herramientas de búsqueda y recuperación.
- Mecanismo: Generación Aumentada por Recuperación (RAG) utilizando documentación proporcionada por el cliente.
- Almacenamiento Vectorial: Base de datos MongoDB utilizada para almacenar embeddings, documentos vectorizados y datos operativos de sesión.

Tareas Clave

El agente actúa como un especialista en información. Procesa consultas en lenguaje natural para proporcionar detalles sobre la flota de vehículos, requisitos de alquiler, paquetes turísticos y políticas de la empresa. Si una consulta queda fuera de su base de conocimientos o requiere intervención humana específica, el agente recopila de forma autónoma los datos de contacto (Nombre, Teléfono, Correo electrónico) y activa una escalación por correo electrónico al equipo de GTN.

Uso Previsto

Usuarios Principales

Clientes externos y turistas que buscan soporte preventa o información general de servicios a través de WhatsApp.

Uso Fuera de Alcance (No Autorizado)

- **Procesamiento de Transacciones:** No autorizado para procesar reservas o reservaciones.
- **Finanzas:** No puede proporcionar precios finales vinculantes ni modificar contratos de facturación.
- **Asesoría:** Prohibido proporcionar asesoramiento contractual.
- **Gestión de Cuentas:** No puede modificar cuentas de usuario existentes ni datos sensibles.
- **Servicios de Emergencia:** Este agente no es para reportar accidentes o emergencias en carretera; los usuarios deben contactar directamente a las líneas de emergencia de GTN.

Datos y Privacidad

Fuentes de Datos

- Manuales Internos de Productos y Servicios 2025-2026.
- Base de datos oficial de preguntas frecuentes (FAQ) internas de GTN.
- Contenido verificado del sitio web de GTN.

Retención de Datos y Seguridad

- Política de PII: No se almacenan datos de identificación personal (PII) de clientes dentro del conjunto de entrenamiento estático del modelo.
- Cifrado: Todos los datos están cifrados en tránsito.
- Guardrails (Protecciones): Se implementa un nodo “Guardrails” en n8n para validar la entrada del usuario y la salida del modelo.
    - Filtros: Escanea PII (contraseñas, llaves, direcciones), toxicidad e intentos de *jailbreak*.
    - Seguridad: Bloquea automáticamente intentos de eludir las medidas de seguridad de la IA.

Consentimiento de Entrenamiento

El modelo y su tubería RAG son privados; los datos procesados no contribuyen a grupos de entrenamiento de terceros ni a las mejoras generales de los modelos de OpenAI.

Cero Entrenamiento con Datos de API:

Por defecto, OpenAI no utiliza los datos enviados a través de su API para entrenar sus modelos fundacionales.

- Privacidad de Entrada: Los *prompts* enviados desde n8n no se añaden al grupo de entrenamiento global.
- Privacidad de Salida: Las respuestas generadas y enviadas a los usuarios de WhatsApp también están excluidas del entrenamiento.

Flujo RAG Privada:
En el entorno n8n, el Flujo RAG es estructuralmente privado:

- Almacenamiento: Los documentos se almacenan en su base de datos MongoDB, no en los servidores de OpenAI.
- Recuperación: Los datos se envían a OpenAI solo como un "Mensaje de Usuario" temporal para esa sesión específica y se eliminan de los registros después de 30 días.

Riesgo y Mitigación (Gobernanza)

Índice de Alucinación

Para calcular el riesgo asociado con una alucinación, aplicamos la Fórmula de Riesgo de IA Bridgeware:

Puntaje de Riesgo (1-100) = [Criticidad + Sensibilidad de Datos] x [Complejidad + Autonomía]

1. Variables y Calificación (Escala 1-10)

Variable	Puntaje	Justificación
Criticidad	4	Moderada. Una alucinación sobre disponibilidad podría causar insatisfacción.
Sensibilidad de Datos	3	Baja-Moderada. El proceso de escalación involucra PII (Email/Teléfono).
Complejidad	5	Moderada. El uso de RAG y MongoDB añade capas donde pueden ocurrir errores de recuperación.
Autonomía	6	Alta. El agente opera 24/7 sin revisión humana previa de cada mensaje.

2. **El Cálculo**

- Paso 1 (Impacto): 4 (Criticidad) + 3 (Sensibilidad) = 7
- Paso 2 (Probabilidad): 5 (Complejidad) + 6 (Autonomía) = 11
- Paso 3 (Resultado Final): 7 x 11 = 77

3**. Actualización de Seguridad y Fiabilidad**

Métrica	Estado	Plan de Implementación
Índice de Alucinación	Pendiente de Verificación	Objetivo: < 2%. Actualmente en fase de "Validación Pre-Despliegue" con un conjunto de datos de prueba (Golden Dataset).
Mitigación de Sesgos	Activo	Filtro de toxicidad en n8n y anclaje estricto a los manuales de GTN (RAG).
Humano en el Bucle	Activo	Escalación automática si el sentimiento es < 0.3 o si hay baja confianza en la recuperación de datos.

Estado de Riesgo Actual: Gestionado (77/100)

Este puntaje de "Riesgo Medio-Alto" nos obliga a mantener una cautela extrema durante la fase de lanzamiento.

Nuestra Garantía de "Seguridad Primero":

- Sin Adivinanzas: Si la IA no está 90% segura de la respuesta en los manuales oficiales, tiene prohibido responder.
- Respaldo Humano Automático: En lugar de dar una respuesta insegura, solicitará sus datos para que un equipo humano de GTN lo contacte.
- Objetivo de Precisión: Trabajamos hacia un entorno de "Cero Errores" (objetivo < 2%).

Consideraciones Éticas

- Transparencia: El agente se identifica explícitamente como "Asistente de IA" al inicio de cada interacción.
- Impacto Ambiental: Al utilizar modelos "mini", minimizamos el consumo de energía y la huella de carbono.