Aplicación de IA Generativa en Consultas de Atención al Cliente

Consultor:

Luis Armando Lazarte Pante

Lima, 2025

Contenido

Detalle metodológico 3

Datos relevantes 4

Arquitectura de la solución 6

Presupuesto 7

Detalle metodológico

Para extraer insights relevantes desde conversaciones reales entre clientes y agentes, se diseñó un enfoque basado en inteligencia artificial que permite construir variables con significado. Las técnicas empleadas combinan modelos de lenguaje de última generación con reglas lingüísticas y estrategias de prompt engineering. A continuación, se describen los principales componentes:

1. Duración en la Comunicación

Se midieron los tiempos de interacción entre el cliente y los distintos roles del equipo (Agente y Supervisor) para capturar indicadores de complejidad, escalamiento o demora en la resolución. Esta variable permite identificar casos con potencial insatisfacción o necesidad de mejora operativa.

2. Análisis de Sentimiento (RoBERTuito - Hugging Face)

Se utilizó el modelo RoBERTuito, especializado en español y entrenado en contextos sociales, para clasificar la polaridad emocional de los mensajes (positivo, negativo, neutral). Su elección responde a la necesidad de trabajar con texto coloquial en español latinoamericano, más representativo del lenguaje real usado por los clientes.

3. Identificación de Puntos Claves (GPT - ChatGPT)

Se implementaron prompts personalizados sobre ChatGPT para identificar tres elementos esenciales de cada interacción:

- Motivo principal del contacto.
- Problema o error reportado.
- Necesidades no satisfechas (explícitas o implícitas).

GPT fue elegido por su capacidad para comprender el contexto conversacional y sintetizar información compleja en frases claras y precisas, superando técnicas tradicionales de extracción de información.

4. Spanglish en los Prompts (Prompt Engineering)

Se aplicó una estrategia híbrida ("Spanglish") en la redacción de los prompts, combinando instrucciones en inglés (idioma de entrenamiento principal de GPT) con contenido en español (idioma del dominio). Esta decisión técnica mejora notablemente la precisión y comprensión del modelo, ya que:

- Las instrucciones en inglés aprovechan la mayor cantidad de datos de entrenamiento de GPT en ese idioma.
- El contenido en español conserva la riqueza contextual del dominio real.

Datos relevantes

1. Problemas de Roaming Internacional

Datos relevantes:

- Las llamadas por roaming internacional tienen la duración promedio más alta: 5.50 minutos.
- Estas llamadas implican diversidad de incidencias: fallos técnicos, configuración, facturación o información.
- Participan múltiples actores (Entel y operadores extranjeros).
- Los agentes carecen de contexto adecuado (país, configuración, historial).
- La información necesaria está dispersa en múltiples sistemas.

Justificación de IA Generativa:

- La herramienta analizaría el problema en tiempo real y sugeriría soluciones probables.
- Generaría resúmenes de llamadas anteriores.
- Consolidaría información relevante para el agente.

Beneficios:

- Reducción del tiempo de llamada.
- Mayor satisfacción del cliente.
- Mejora en la formación de agentes.

2. Problemas de Facturación y Servicios

Datos relevantes:

- Son los motivos de contacto más frecuentes.
- Generan alto sentimiento negativo.
- En especial, "Discrepancias en facturación" requieren intervención de supervisores.

Justificación de IA Generativa:

- Crear asistentes que analicen facturas y detecten discrepancias.

- Proporcionar respuestas empáticas y precisas.
- Desarrollar agentes de autoservicio para problemas comunes.

Beneficios:

- Reducción del volumen de llamadas.
- Disminución del tiempo de resolución.
- Menor necesidad de escalamiento.

3. **Problemas de Alto Impacto pero Baja Frecuencia**

Datos relevantes:

- Ej: "producto defectuoso", "retraso en la entrega", "falta de servicio".
- Tienen 100% de sentimiento negativo.

Justificación de IA Generativa:

- Asistir al agente con protocolos de atención sensibles.
- Generar alertas sobre casos delicados.
- Monitoreo proactivo de entregas y productos.

Beneficios:

- Mejora de la percepción del cliente.
- Prevención de quejas y mala reputación.

4. **Problema identificado**

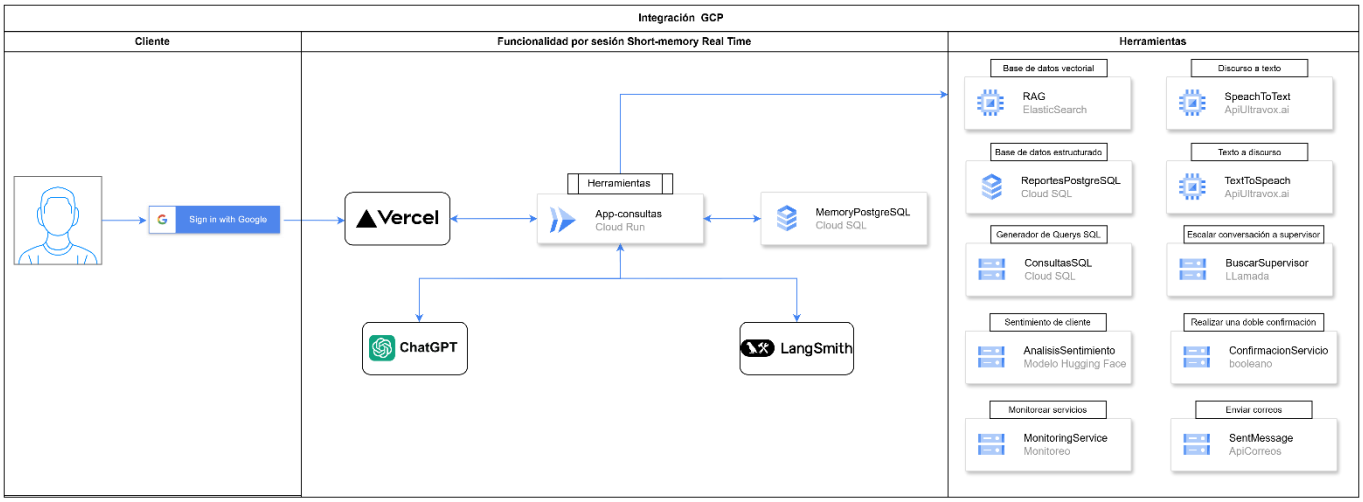
Tipo de problema	**Solución recomendada**
Facturación y servicios frecuentes	Chatbots, FAQs dinámicas, agentes de autoservicio
Roaming internacional	IA asistente para agentes con contexto, soluciones y resumen
Soporte de decodificadores de TV	Guías interactivas generadas por IA
Discrepancias en facturación	IA para análisis y resolución empática

5. **Necesidades Identificadas y Soluciones**

Necesidad del Cliente	**Solución recomendada**
Solución rápida de internet	Diagnóstico automático y sugerencias al cliente
Fiabilidad en facturación	Validación automática y explicación clara
Comunicación clara y proactiva	Notificaciones proactivas

Consentimiento garantizado	Sistema de doble confirmación automático

Arquitectura de la solución



Integración GCP:

Este bloque general representa cómo se orquesta el flujo entre el cliente, los servicios en la nube.

Cliente:

- Sign in with Google: Permite al usuario autenticarse mediante su cuenta de Google.

Funcionalidad por sesión (Short-memory Real Time):

- ChatGPT: Interfaz de interacción donde el usuario realiza consultas en lenguaje natural.
- Vercel: Plataforma para desplegar el frontend de la aplicación.
- App-Consultas - Cloud Run: Microservicio en GCP que maneja la lógica para procesar la consulta del usuario y coordinar el flujo hacia los componentes internos.
- Memory/PostgreSQL (Cloud SQL): Almacena el historial breve (short memory) de las interacciones para cada sesión.
- ElasticSearch (Compute Engine): Motor de búsqueda que permite realizar consultas semánticas eficientes sobre los documentos indexados usando los metadatos.

Presupuesto

Componente	Características	Precio
PostgreSQL	- 128 CPU virtual(es) - 250 GB - IP privada - 864 de RAM	$ 33.53 por hora
VM	- Útil para inferencia y despliegue. - M4	$ 2.21 por hora
Buckets		$0.026 por GB al mes
Cloud run		$0.000024 por vCPU-segundo $0.00000250 por GiB-segundo $0.40 por millón
Api ultravox.ai	- Llamadas de 30 segundos ilimitadas - Hasta 100 llamadas simultáneas con prioridad.	$ 1000 por mes
LangSmith	- Puedes registrar más ejecuciones simultáneas.	$ 39 por mes
GPT-4o	- Multimodalidad - Más eficiente y barato que GPT-4 - Soporte multilingüe fluido - Comprensión contextual exelente	$2.50 por millón de tokens de entrada $10.00 por millón de tokens de salida