BRIDGEWARE
Technologies

AI Model Card

Project: Customer Support Intelligent Agent v1.0
June 2025

Developer:
Bridgeware Technologies

This model card outlines the operational and ethical framework for the GTN WhatsApp Intelligent Agent v.1.0. This agent serves as a frontline AI Assistant for car rental and tourism inquiries, bridging the gap between automated information retrieval and personalized human support.

Model Details

Core Architecture

- **Platform:** n8n workflow automation platform.
- **LLM Engine:** GPT-4.1-mini via OpenAI API.
- **Communication:** Meta WhatsApp Business API.
- **Integrations:** Gmail via Google Cloud API (for lead escalation).

Training Type

- **Framework:** Model Context Protocol (MCP) used to expose search and retrieval tools.
- **Mechanism:** Retrieval-Augmented Generation (RAG) utilizing client-provided documentation.
- **Vector Store:** MongoDB Database used to store embeddings, vectorized documents, and operational session data.

Key Tasks

The agent acts as a **specialized information specialist**. It processes natural language queries to provide details on vehicle fleets, rental requirements, tourism packages, and company policies. If a query falls outside its knowledge base or requires specific human intervention, the agent autonomously collects contact details (Name, Phone, Email) and triggers an email escalation to the GTN team.

Intended Use

Primary Users

External customers and tourists seeking pre-sales support or general service information via WhatsApp.

Out-of-Scope Use

- **Transaction Processing:** Not authorized to process bookings or reservations.
- **Financials:** Cannot provide final binding pricing or modify billing contracts.
- **Advisory:** Prohibited from providing contractual advice.
- **Account Management:** Cannot modify existing user accounts or sensitive data.

• **Emergency Services:** This agent is not for reporting accidents or roadside emergencies; users should contact GTN emergency lines directly.

Data & Privacy

Data Sources

- 2025-2026 Internal Product and Services Manuals.
- Official GTN Internal FAQ database.
- Scraped and verified GTN Website Content.

Data Retention & Security

- **PII Policy:** No customer PII is stored within the model's static training set.
- **Encryption:** All data is encrypted in transit.
- **Guardrails:** An n8n Guardrails node is implemented to validate user input and model output.
    - **Filters:** Scans for PII (passwords, keys, addresses), toxicity, and jailbreak attempts.
    - **Safety:** Automatically blocks attempts to bypass AI safety measures.

Training Consent

The model and its RAG pipeline are private; data processed does not contribute to public third-party training pools or OpenAI's general model improvements.

Zero Training on API Data

By default, OpenAI **does not** use data submitted via their API to train their foundational models or improve their general services. This is a contractual guarantee for all API customers.

- **Input Privacy:** The prompts sent from n8n "OpenAI Node" are not added to the global training pool.
- **Output Privacy:** The responses generated and sent back to WhatsApp users are also excluded from training.

Private RAG Pipeline

In an n8n environment, **RAG (Retrieval-Augmented Generation)** pipeline is structurally private:

- **Storage:** Documents and embeddings are stored in **your MongoDB database**, not OpenAI's servers.
- **Retrieval:** When n8n fetches a document snippet to provide context to the model, that data is only sent to OpenAI as a temporary "User Message" for that specific session. It is treated as ephemeral data and is typically deleted from OpenAI's logs after 30 days (standard retention for abuse monitoring).

n8n Node Data Handling

When using the OpenAI node in n8n:

- **Credential Security:** n8n stores your OpenAI API key in an encrypted format. It is never sent to OpenAI as part of the prompt.

- **Execution Logs:** If you are self-hosting n8n, you have 100% control over where the execution data is stored. If you use n8n Cloud, they follow a **No-Training** policy as well, ensuring your workflow logic and data aren't used to improve their own internal AI assistants.

Risk & Mitigation (Governance)

Hallucination Rate

To calculate the risk associated with a **Hallucination** in your GTN WhatsApp Agent, we apply the **Bridgeware AI Risk Formula**. This formula moves beyond simple probability by looking at the environmental factors (Criticality and Data Sensitivity) and the operational factors (Complexity and Autonomy).

Based on this Model Card—where the agent provides info but **cannot** perform bookings—here is the risk profile for a hallucination event.

1. Variables & Scoring (Scale 1-10)

Variable	Score	Justification
Criticality	**4**	Moderate. While the agent doesn't handle payments, a hallucination about vehicle availability or tourism requirements could lead to customer dissatisfaction or travel disruptions.
Data Sensitivity	**3**	Low-Moderate. The model uses public manual data, but the escalation process involves PII (Email/Phone). Hallucinating a privacy breach or leaking a prompt is a minor risk.
Complexity	**5**	Moderate. The use of RAG and MongoDB adds layers where retrieval errors (fetching the wrong info) can occur.
Autonomy	**6**	High. The agent operates 24/7 on WhatsApp without a human reviewing every message before it is sent.

2. The Calculation

Risk Score (1-100) = [Criticality + Data Sensitivity] x [Complexity + Autonomy]

Step 1: Calculate Impact

4 (Criticality) + 3 (Data Sensitivity) = 7

Step 2: Calculate Probability

5 (Complexity) + 6 (Autonomy) = 11

Step 3: Final Risk Score

7 x 11 = 77

3. Safety & Reliability Update

Metric	Status	Implementation Plan
Hallucination Rate	**Pending Verification**	Target: < 2%. System is currently in the "Pre-Deployment Validation" phase. A "Golden Dataset" of 100+ GTN-specific support scenarios is being curated for automated RAGAS (Retrieval-Augmented Generation Assessment) testing.
Bias Mitigation	**Active**	Inputs are passed through an n8n-integrated toxicity filter. Context-injection (RAG) is used to strictly anchor responses to GTN Manuals, reducing "creative" model drift.
Human-in-the-Loop	**Active**	Escalation Trigger: Any session with a Sentiment Score < 0.3 or a "Low-Confidence" retrieval flag from MongoDB is automatically routed to a human GTN agent via Gmail.

Current Risk Status: Managed (77/100)

To ensure your experience is safe and accurate, we use a standard AI Risk Formula to grade our assistant. We currently give the system a Safety Score of 77/100, which falls into the "High-Medium Managed" category. This rating helps us stay extra cautious while the assistant is in its early launch phase.

How We Calculated This Score:

Factor	Score	Simple Explanation
Importance of Task	**4/10**	The AI provides travel info but cannot take payments or book cars, which limits the risk of financial errors.
Data Privacy	**3/10**	The AI uses public brochures. It only asks for your contact info if it gets stuck, so it can pass your request to a real person.
System Complexity	**5/10**	Because the AI looks up information from our live manuals (RAG), there is a small chance it could misread a detail while we are still fine-tuning it.
Independence	**6/10**	The AI works 24/7 on WhatsApp without a human reading every single reply, which is why we have strict "No Guessing" rules in place.

Our "Safety-First" Guarantee

Because we are still in the final stages of "stress-testing" this assistant for perfect accuracy, we have programmed it with a Strict Honesty Policy:

- No Guessing: If the AI is not absolutely sure it has found the right answer in our official GTN manuals, it is forbidden from answering.
- Automatic Human Backup: Instead of giving customers an unsure answer, the AI will immediately ask for name, email and phone number so a GTN Team Member can personally call or email customer with the correct information.
- Accuracy Goal: We are working toward a "Zero-Error" environment, with a target of less than 2% of total conversations needing a correction.

How we are currently mitigating this:

- **Guardrails Node:** Validates output to ensure the hallucination doesn't include fake PII or system keys.
- **RAG Architecture:** Limits the "creativity" of GPT-4.1-mini by forcing it to look at the MongoDB documentation first.
- **Sentiment Escalation:** If a hallucination causes a customer to become frustrated (Score < 0.3), a human takes over, capping the "Impact" duration.

Bias Mitigation

All inputs pass through a dedicated toxicity and bias filter. The agent is programmed to remain neutral and helpful, avoiding subjective comparisons or culturally insensitive suggestions.

Human-in-the-Loop

The system monitors user sentiment in real-time. Any session yielding a **Sentiment Score below 0.3** (indicating high frustration or anger) triggers an automatic "Hand-off" flag, notifying the GTN team immediately for manual intervention.

Ethical Considerations

- **Transparency:** The agent explicitly identifies as an **"AI Assistant"** at the start of every interaction to manage user expectations.
- **Environmental Impact:** By utilizing "mini" model variants (GPT-4.1-mini), the project minimizes the tokens processed and the associated compute energy, favoring a lower carbon footprint for high-volume support tasks.