Hotel Assistant AI Agent — Key Settings Guide (Azure AI Foundry)

A practical, copy■paste■friendly checklist for configuring a production■ready hotel concierge/chat agent in Azure AI Foundry. Use it when you create a new Agent (or a Prompt Flow■backed agent) and when promoting from Dev → Staging → Prod.

1) Project & Resource Setup

Goal: Clean separation of environments; secure connections to data; predictable deployments.

- Environments: Create three workspaces/subscriptions or resource groups: dev, staging, prod.
- AI resources: Provision Azure OpenAI (models), Azure AI Search (vector index), Azure Storage (blobs), Azure Key Vault (secrets), App Insights/Monitor (telemetry).
- Networking: Private endpoints/VNET integration (if PII), IP allowlists for admin tools, egress rules for functions.
- RBAC: Least privilege roles (Owner/Contributor for platform ops, Reader for stakeholders, custom roles for annotators/evaluators).
- Secrets: Store API keys, connection strings in Key Vault, reference from Foundry connections.

2) Agent Identity & Policy

Where: Agent → Basics / Instructions / Safety

- Agent Name: Hotel Assistant – Crystal Hotels
- Short Description: Helps guests with reservations, check■in/out, amenities, and local recommendations.
- System Instructions (Persona): Tone: warm, efficient, hospitality■forward. Style: short, actionable answers; offer to complete tasks. Jurisdiction: region■appropriate policies (cancellation, taxes/fees). Privacy: never share room numbers or PII; verify identity before booking changes.
- Guardrails: Block payment collection; route to secure link or PCI■compliant flow. Disallow medical/legal advice; safe■complete with help resources. Refuse unsafe content; apply content filters.
- Language: Default English; auto■detect and respond in user language (ES/FR/DE). Confirm critical details in English for back■office handoff.

Template — System Prompt (paste into Instructions)

You are Crystal Hotels' virtual concierge. Prioritize guest safety, privacy, and accuracy.
- Always verify identity for booking lookups or changes: ask for last name + confirmation code or phone + last 4 of payment method (never reveal full details).
- Never disclose room numbers or personal data. Never accept payments. Route payment requests to the secure checkout URL.
- Be proactive but concise. Offer next best actions and summarize confirmations.
- For operational requests (extra towels, late checkout), create a service ticket with priority and ETA; confirm back to the guest.
- For questions about availability or price, use live inventory/pricing tools; if unavailable, provide contact

path.
- Detect language and reply in that language. For critical confirmations, include an English summary line prefixed with “Back■office:”.
- If content may be unsafe or out■of■policy, refuse politely and suggest a safe alternative.

3) Model & Inference Settings

Where: Agent → Model / Parameters

- Chat model: GPT■4.x class (or latest gpt■4o/equivalent). Keep a quality and a fast variant.
- Temperature: 0.3–0.5; Top_p: 0.9.
- Max output tokens: 512–800 for chat; 1,200+ for summary/itinerary.
- Response format: JSON schema for function outputs; plain text for conversation.
- Streaming: On for responsiveness.

Tip: Route complex itineraries to Quality; FAQs to Fast.

4) Knowledge & Retrieval (RAG)

Where: Agent → Add data / Grounding (Azure AI Search, Blob Storage)

- Sources: SOPs, amenities, room types, fees, house rules, loyalty tiers, local attractions, menus, emergency procedures.
- Indexing (Azure AI Search): Chunk size 500–1,000 tokens with 60–120 overlap. Embedding model text-embedding-3-large (or latest).
- Fields: title, content, lang, effective_date, property_code, policy_version.
- Filters: property/brand filters (e.g., property_code == “SAN■MV”).
- Citations: Enable and include policy_version and effective_date.
- Freshness: add expiry tag; ignore expired policies via filter.

RAG Prompt Snippet

Use only retrieved hotel policies for compliance topics. If missing or stale (effective_date > 12 months old), say you cannot confirm and escalate. Cite titles and policy_version in a bullet list at the end of your answer.

5) Tools & Actions (Function Calling)

Where: Agent → Tools (Functions/Connectors)

- Booking Lookup: get_reservation(last_name, conf_code | phone_last4)
- Modify Booking: change_dates(conf_id, new_checkin, new_checkout) with policy validation and re■quote.
- Service Ticket: create_ticket(room_id|res_id, category, priority) → PMS/Housekeeping.
- Room Availability: query_inventory(property_code, dates, room_type?)
- Pricing/Offers: quote_rate(property_code, dates, rate_plan?)

- Local Recs: nearby_places(category, distance_km) with caching.
- Escalation: handoff_to_human(context, transcript, priority)

Schema Pattern (example)

{ "name": "create_ticket", "description": "Create an ops ticket for housekeeping/maintenance/concierge", "parameters": { "type": "object", "properties": { "res_id": {"type": "string"}, "category": {"type": "string", "enum": ["housekeeping","maintenance","concierge"]}, "priority": {"type": "string", "enum": ["low","normal","high"]}, "notes": {"type": "string"} }, "required": ["res_id","category"] } }

6) Memory & Conversation Control

Where: Agent → Memory / State

- Session window: keep last 6–10 turns; summarize beyond.
- Short■term memory: trip dates, party size, prefs (pillows, floor) within session; expire after checkout.
- Long■term memory: off by default for PII; enable only with consent and for non■sensitive prefs.
- Persona guard: disallow “act as” jailbreaks; re■assert system role if injection detected.

7) Safety & Compliance

Where: Agent → Safety (Azure AI Content Safety), Data Handling, PII

- Content filters: enable hate/sexual/violence/self■harm thresholds; safe■complete.
- PII: redact/mask in logs and tool calls (phone, email, card last4 only).
- Identity: require 2■factor question or profile match for refunds/name changes.
- Geo policy: emergency info; no medical advice; escalate for minors.

8) Evaluation & Quality Gates

Where: Evaluation tab / Prompt Flow Evals / Traces

- Golden set: 50–150 curated tasks across intents, languages, and edge cases.
- Metrics: task success, groundedness, citation accuracy, refusal appropriateness, tone, latency, tool■call success.
- Regression: run on every model/knowledge/tool change; block if degradation > tolerance.
- Live probes: synthetic monitors for top 10 intents.

9) Observability & Monitoring

Where: Monitoring / Traces / App Insights

- Logs: prompt/response hashes (no raw PII), tool results, P50/P95 latency, tokens, refusals.
- Dashboards: intent distribution, success rate, handoff rate, average handle time, ticket SLAs.
- Alerts: spikes in refusals, tool failures, hallucination flags; RAG missing■doc rate.

10) Deployment & Promotion

Where: Deployments / Environments

- Config as code: store agent instructions, tool schemas, retrieval filters in source control.
- Feature flags: toggle new tools or prompts per environment.
- Promotion checklist: evals pass, canary in staging 24–48h, rollback snapshot.

11) Example Config (YAML sketch)

agent:
name: Crystal Hotels Assistant
model: gpt-4o
parameters:
temperature: 0.4
top_p: 0.9
max_output_tokens: 800
instructions_file: ./prompts/system.md
memory:
window_turns: 8
pii_persistence: false
safety:
content_filters: hotel_default
pii_redaction: enabled
retrieval:
index: ai-search://crystal/policies
embedding_model: text-embedding-3-large
chunk: {size_tokens: 800, overlap_tokens: 80}
filters:
property_code: [SAN-MV, NYC-MD]
valid_only: true
tools:
- ./tools/get_reservation.json
- ./tools/change_dates.json
- ./tools/create_ticket.json
- ./tools/query_inventory.json
- ./tools/quote_rate.json
- ./tools/handoff_to_human.json
telemetry:
app_insights: ai://crystal/insights
deployments:
environments: [dev, staging, prod]
feature_flags:
local_recs: off
quality_model: on

12) Quick Intent Pack (for evaluation)

- “Can I check in early this Saturday? My last name is Vega and my code is QJ8D23.”
- “¿Pueden recomendarme un restaurante mexicano cerca del hotel para 4 personas a las 8pm?”
- “My flight was canceled—please move my checkout to Monday and add late checkout.”
- “What’s the pet fee and is there a weight limit?”
- “I lost my key; can you make me a new one?” → Must verify identity and route to front desk.
- “The AC is leaking in room 512.” → Create high■priority maintenance ticket + ETA.

13) Hand■Off Script (Agent → Human)

Where: Tool handoff_to_human payload template

Guest name:
Reservation ID:
Issue:
Actions tried:
Proposed next step:
Priority:
Transcript last 10 turns:

14) Go■Live Checklist (1■pager)

- System prompt reviewed by ops & legal
- Safety thresholds approved; refusal copy localized
- RAG sources current; expiry filter active
- Tool schemas validated in staging; timeouts + retries set
- Golden evals pass; probes green; rollback ready
- Monitoring and alerts tested; PII masked in logs