Part 5: Agentic AI Lead Qualification Design

1. Objective

Automate first-call lead qualification to (a) reduce speed-to-contact from hours to minutes, (b) standardize discovery questions, (c) enrich leads with structured intent + disqualification reasons, and (d) free sales reps to focus on high-intent leads. The AI voice agent places/receives calls to new leads, conducts a scripted yet adaptive qualification conversation, updates `lead_status`, captures structured fields, and schedules human follow-ups when needed.

2. High-Level Architecture

```
Lead Ingestion (campaign_leads.csv) → Event Bus (NewLead) → Scheduling &
Prioritization Service
    → Call Orchestrator → Telephony/RTC Layer (e.g., Twilio Voice, SIP) ⇔ Real-
time AI Runtime
        (ASR + Streaming LLM + TTS + Dialog State Manager + Tool Functions)
            ⇔ CRM/Data Services (Lead Status, Notes, Metadata, Consent Logs)
                → Analytics & Monitoring (Quality, Latency, Outcomes)
                    → Feedback Loop (Fine-tuning, Prompt Optimization)
```

Mermaid Flowchart (Operational Call Lifecycle)

```
flowchart LR
    A[New Lead Event] --> B{Eligible Time Window?}
    B -->|No| W[Wait / Requeue]
    B -->|Yes| C[Scheduler & Priority Queue]
    C --> D[Call Orchestrator]
    D --> E[Telephony Provider PSTN/SIP]
    E --> F(Bi-directional Audio Stream)
    F --> G[ASR Engine]
    F --> H[TTS: ElevenLabs]
    G --> I[Dialog State Manager]
    I --> J[LLM Reasoner]
    J --> I
    H --> E
    I --> K{Slots Complete or Disqualified?}
    K -->|No| F
    K -->|Yes| L[Persist Session & Summary]
    L --> M[Update Lead Status]
    M --> N[Analytics & Monitoring]
    N --> O[Feedback Loop / Fine-Tuning]
    O --> J
```

ElevenLabs TTS Rationale

Chosen initially for: (1) high-quality natural prosody suited to sales qualification, (2) low latency streaming endpoints, (3) multi-voice & language support (English/Arabic roadmap), (4) adjustable stability/similarity parameters enabling tone experimentation, (5) straightforward API integration allowing later swap if cost profile changes.

Latency Target: < 300ms synthesis per utterance for snappy turns (overall ASR→LLM→TTS cycle < 1.2s).

Fallback Strategy: If ElevenLabs API latency > threshold or error rate spikes, switch to cached utterance templates or secondary TTS provider (Azure Neural) while logging degradation events.

ElevenLabs Integration Snippet (Python)

```
import requests

ELEVENLABS_API_KEY = "YOUR_API_KEY"  # store securely (env var / secret manager)
DEFAULT_VOICE_ID = "EXAVITQu4vr4xnSDxMaL"  # example voice

def synthesize_tts(text: str,
                   voice_id: str = DEFAULT_VOICE_ID,
                   model_id: str = "eleven_monolingual_v1",
                   stability: float = 0.3,
                   similarity_boost: float = 0.8) -> bytes:
    """Return MP3 audio bytes for the given text using ElevenLabs."""
    url = f"https://api.elevenlabs.io/v1/text-to-speech/{voice_id}"
    headers = {
        "xi-api-key": ELEVENLABS_API_KEY,
        "Content-Type": "application/json"
    }
    payload = {
        "text": text,
        "model_id": model_id,
        "voice_settings": {
            "stability": stability,
            "similarity_boost": similarity_boost
        }
    }
    resp = requests.post(url, json=payload, headers=headers, timeout=10)
    resp.raise_for_status()
    return resp.content  # MP3 bytes

# Usage inside dialog loop (simplified):
def speak_response(call_id: str, response_text: str):
    audio_mp3 = synthesize_tts(response_text)
    # convert MP3 -> PCM if telephony requires raw RTP, then stream
    stream_audio_to_call(call_id, audio_mp3)
```

Future TTS Optimization

- Pre-generate frequent short utterances (greetings, acknowledgements) and cache.
- Batch multi-sentence synthesis to reduce per-request overhead when appropriate.

- Monitor mean, p95 synthesis latency; trigger voice model downgrade if costs grow.

Core Components

Component	Responsibility	Candidate Tech
Scheduling Engine	Optimal call time selection; retries; throttle	Python service + Redis (priority queue)
Call Orchestrator	Lifecycle: dial, connect, hand-off, terminate	FastAPI microservice
Telephony Layer	PSTN/SIP, DTMF, call recording, compliance	Twilio / Plivo / WebRTC media server
ASR (Speech-to-Text)	Low-latency streaming transcription	OpenAI Realtime / Whisper.cpp / DeepGram
TTS (Text-to-Speech)	Natural voice; emotional prosody	PlayHT / ElevenLabs / Azure Neural
Dialog Manager	State machine + policy + memory slots	Custom Python + Redis session state
LLM Reasoner	Intent extraction, next question selection	GPT-4.1/GPT-5* (Hosted)
Structured Extractor	Map free text → schema fields	LLM JSON mode + validation layer
Compliance & Consent	GDPR/CCPA consent phrase logging	Policy module + timestamped audit DB
Data Warehouse	Persist interactions, metrics	PostgreSQL (OLTP) + Snowflake (analytics)
Monitoring	Quality, latency, error alerts	Prometheus + Grafana + Sentry
Evaluation Pipeline	Scoring transcripts vs ground truth	Python batch jobs + ML (classification)

(*When asked about model use in product positioning, vendor marketing can refer to GPT-5 readiness, but exact version abstracted to users.)

Data Flow Details

1. New lead ingested → event published `lead.created` with campaign_id.
2. Scheduler checks allowable call window (time zone, local calling regulations) + priority (channel, spend, lead source scoring). If immediate window open: enqueue call.
3. Orchestrator initiates outbound call through telephony provider; upon connect starts bi-directional audio streaming to AI runtime.
4. Streaming ASR produces partial transcripts; Dialog Manager updates slot-filling state; LLM decides next utterance; TTS renders audio; latency target < 1.2s turnaround.
5. At end-of-call, structured summary JSON persisted: `{qualification_status, budget_range, timeline, property_type, disqualification_reason, sentiment_score}`.
6. `lead_status_changes.csv` appended with new status events (e.g., QUALIFIED, NOT_QUALIFIED, CALL_AGAIN, WRONG_NUMBER).

7. Analytics aggregates per campaign: contact rate, qualification rate, average call duration, cost per qualified contact.
8. Feedback loop compares AI-decided qualification vs later human override; disagreement examples curated for prompt refinement or model fine-tuning.

3. Functional Scope

MVP Scope (Phase 1)

- Outbound calls within allowed hours (local time inference from phone country code).
- Scripted qualification for real estate leads: confirm interest, budget, timeline, property type, financing readiness.
- Basic disqualification reasons: Wrong Number, Already Bought, Low Budget, Not Interested.
- Update status + structured fields + call recording.
- Dashboard view: AI Contact Attempts, Success Rate, Average Handle Time (AHT), Qualification Rate vs Human baseline.

Phase 2 Enhancements

- Inbound call handling (callback requests, missed human calls).
- Calendar integration: schedule human follow-up meeting / site visit.
- Multi-language (Arabic/English) auto-detection + bilingual script.
- Sentiment & urgency scoring impacting priority queue.
- Adaptive script experimentation (A/B test variants).

Phase 3 (Intelligent Optimization)

- Automatic dynamic rephrasing using real-time user sentiment.
- Predictive callback timing (model: probability lead answers next hour).
- Portfolio-level qualification forecasting by campaign.
- Auto-handoff to human rep mid-call if high deal value or complex objection.

4. User Workflows & Interaction Points

1. Enable Feature: User toggles “AI Qualification” per campaign with guardrails (max daily AI calls, time windows).
2. Script Configuration: Admin tweaks question ordering, required fields, fallback answers, disclaimers.
3. Monitoring: Real-time dashboard card “AI Agent Performance” with KPIs.
4. Review & Override: Sales reps view transcript + structured summary; can adjust status (feedback captured).
5. Escalations: If agent marks “HIGH_INTEREST” or identifies large budget, system triggers task for senior rep.
6. Compliance Audit: Export consent log + recordings for selected date range.

5. Conversation Design

Slot Schema

```
slots = {
  "interest_level": ["HIGH", "MEDIUM", "LOW"],
  "budget_range": {"min": int, "max": int, "currency": "USD"},
  "property_type": ["Apartment", "Villa", "Commercial", "Other"],
  "purchase_timeline": ["<30d", "30-90d", ">90d"],
  "financing_ready": bool,
  "lead_intent_notes": str,
  "disqualification_reason": str|null
}
```

Sample Flow (Qualified Path)

1. Greeting & Consent: “Hi, this is the Leadsmart virtual assistant calling about your interest in Mivida Gardens. Is now a good time?”
2. Confirmation: “Great. To match you with the right options, may I confirm the approximate budget range you’re considering?”
3. Timeline Probe: “When are you hoping to finalize your purchase?”
4. Financing: “Do you already have financing pre-approved or will you be exploring that now?”
5. Property Type: “Are you focused on apartments, villas, or open to options?”
6. Qualification Decision (internal) → If criteria met: mark QUALIFIED.
7. Escalation Offer: “Would you like me to schedule a call with a property specialist tomorrow?”
8. Closing: “Thank you. We’ll follow up shortly. Have a great day.”

Disqualification Examples

- Wrong Number: Early mismatch → “I apologize. I will update our records—no further calls.” Status WRONG_NUMBER.
- Low Budget: Budget < internal threshold vs project price → collect reason; status LOW_BUDGET.
- Already Purchased: “Congratulations. I’ll ensure we stop further outreach.” Status ALREADY_BOUGHT.

Objection Handling (Branching)

Objection	AI Strategy	Follow-up
“Just browsing”	Light nurture, ask timeline	Set CALL_AGAIN, schedule reminder
“Busy now”	Offer callback window	CALL_AGAIN with time slot
“Send info”	Collect email confirm	WHATSAPP/EMAIL follow-up task

System Prompt (Skeleton)

```
SYSTEM:
You are Leadsmart Voice Qualification Agent. Goal: courteously collect slots:
budget_range, purchase_timeline, property_type, financing_ready, interest_level.
NEVER fabricate. After each user utterance, update internal state. When all
mandatory slots filled OR clear disqualification, summarize JSON:
{qualification_status, budget_range, timeline, property_type, financing_ready,
interest_level, disqualification_reason|null}
```

```
Constraints: keep responses ≤18 words, natural tone, no jargon, confirm consent at
start. Avoid repeating answered slots. If user refuses, politely close and mark
disqualification_reason.
```

Tool Invocation Examples

```
tool.schedule_followup({"lead_id":123,"datetime":"2025-11-16T10:30:00+02:00"})
tool.update_status({"lead_id":123,"status":"QUALIFIED"})
tool.persist_slots({...})
```

6. Data Model Extensions

Add table `agent_call_sessions`:

```
id, lead_id, campaign_id, started_at, ended_at, duration_sec,
status_before, status_after, transcript_path, recording_path,
qualification_status, slots_json, escalation_flag, sentiment_avg,
cost_usd, model_version
```

Add table `agent_feedback_overrides` for human corrections.

7. KPI & Evaluation Metrics

Metric	Definition	Target (Phase 2)
Contact Rate	Successful connections / attempted calls	>65%
Qualification Rate	QUALIFIED / reached leads	Within ±5% of human baseline
Avg Handle Time	Mean call length (sec)	< 240s
Speed-to-Contact	Median time lead created → first call	< 5 minutes
Override Rate	Human status changes / AI qualified leads	< 15%
Cost per Qualified Contact	Total agent ops cost / # qualified	< 60% of manual cost
Sentiment Mismatch	(Human review negative) / AI high interest	< 10%
Latency Turnaround	Avg ASR→TTS cycle	<1.2s

8. ROI & Cost-Benefit Model

Variables

```
H = average human dial attempts per new lead (e.g., 3)
C_h = human cost per hour (e.g., $25)
T_call = average human call handling time (minutes, e.g., 6)
```

```
T_admin = post-call admin time (minutes, e.g., 2)
L = leads per month (e.g., 5,000)
Conv_q = proportion of leads qualified (0.12 baseline)
Rev_per_deal = average net revenue per closed deal ($3,000)
Close_rate_q = close probability of qualified lead (0.25)

AI Costs per minute: telephony $0.012, ASR $0.01, TTS $0.008, LLM tokens ~$0.02 → blended $0.05/min
Mean AI call duration = 4 min
Retries (no answer) cost minimal (<15 sec greeting) assumed 0.5 min average additional.
```

Human Monthly Cost

```
Human_minutes = L * (T_call + T_admin)
Human_hours = Human_minutes / 60
Human_cost = Human_hours * C_h
```

AI Monthly Cost

```
AI_minutes = L * (4.0 + 0.5)  # include retries
AI_cost = AI_minutes * 0.05
```

Uplift & Net Benefit

Speed-to-contact improvement can raise Conv_q (e.g., +2 percentage points). Additional qualified leads:

```
Δqualified = L * (Conv_q_new - Conv_q_baseline)
Incremental_revenue = Δqualified * Close_rate_q * Rev_per_deal
Net_benefit = (Human_cost - AI_cost) + Incremental_revenue
ROI = Net_benefit / AI_cost
```

Illustrative numbers (substitute real data) show potential ROI > 3× within 3–6 months.

Sensitivity Analysis Dimensions

- Lead volume variance ±30%.
- ASR pricing renegotiation (-20%).
- Qualification uplift scenarios: +1pp / +2pp / +5pp.
- Call duration creep risk (+1 min) impact.

9. Build vs Buy Analysis

Option	Pros	Cons	Est. Time	Strategic Fit

Option	Pros	Cons	Est. Time	Strategic Fit
Full Custom (Twilio + Whisper + GPT)	Control, extensibility, data ownership	Higher initial engineering + maintenance	10–14 weeks MVP	High (core differentiator)
Vendor Voice AI (e.g., Talkdesk AI)	Faster deploy, enterprise features	Rigid scripting, costly per seat/minute	4–6 weeks integration	Medium
Hybrid (Custom dialog + 3rd-party ASR/TTS)	Balance speed & control	Latency coordination	8–10 weeks	High
Outsource BPO + simple analytics	Immediate scale	No automation; variable quality	2 weeks	Low

Decision Criteria

1. Differentiation vs commodity tele-qualification.
2. Data privacy & ability to fine-tune on domain transcripts.
3. Multi-language + domain adaptation needs.
4. Long-term variable cost trajectory (negotiating infrastructure vs vendor lock-in).
5. Flexibility to embed real-time campaign intelligence (e.g., dynamic pitch from insights).

Preliminary Recommendation: Hybrid build (custom orchestration + LLM reasoning; outsource low-level audio to reliable managed APIs initially; later optimize cost by migrating ASR/TTS in-house if call volume > 100k/month).

10. Assumptions

Category	Assumption	Validation Path
Regulatory	Consent phrase sufficient for jurisdiction	Legal review (jurisdictions served)
Data	Phone numbers include country code reliably	Audit sample of recent leads
Volume	≥ 3,000 leads/month justifies automation	Historical pipeline analysis
Script	Core qualification questions stable across campaigns	Stakeholder workshop
Languages	English + Arabic cover 90% of leads	Lead language distribution stats
Infrastructure	Existing ETL and warehouse available	Review current stack (Part 4 artifacts)
Sales Workflow	Reps adopt transcript review daily	Pilot training & usage tracking
Budget Data	Budget ranges meaningful qualifiers	Compare with property price bins

11. Stakeholder Questions Before Commitment

1. What is current average speed-to-contact and its correlation with qualification? (Need baseline regression.)
2. What % of leads never receive any call attempts? (Uncontacted opportunity size.)
3. Are there legal constraints on automated voice outreach in target markets? (TCPA, GDPR, local telemarketing laws.)
4. How critical is multilingual support at launch?
5. Do we need inbound handling (return calls) in MVP or acceptable for Phase 2?
6. What are priority disqualification codes for revenue reporting?
7. Are reps comfortable with AI scheduling follow-ups directly on calendars?
8. SLA expectations: acceptable average call latency? Max daily failure rate?
9. How will success be judged after 30/60/90 days? (Define adoption & revenue KPIs.)
10. Are there brand tone guidelines the AI must follow verbatim?
11. Data retention policy for recordings & transcripts?
12. Do campaigns vary script materially (e.g., commercial vs residential)? Need template system.

12. Risk & Mitigation

Risk	Impact	Mitigation
ASR errors (accent, noise)	Misqualification	Multi-engine fallback; confidence thresholds; escalate unclear cases
Regulatory non-compliance	Fines / reputational	Jurisdiction-based consent module; legal audit; do-not-call list sync
User mistrust / low adoption	Poor ROI	Transparent transcripts, opt-out controls, early pilot champions
Latency >1.5s	Unnatural conversation	Pre-warm model, streaming inference, reduce prompt size
Cost overrun (token usage)	Negative margin	JSON extraction mode, compress prompts, caching for common utterances
Script stagnation	Lower conversion	Continuous A/B test + outcome analytics
Data drift (new lead types)	Model degradation	Quarterly fine-tune; drift detection on slot fill error rate
Language expansion	Missed leads	Modular NLU layer with pluggable multilingual models

13. MVP Roadmap (Indicative)

Week	Milestone
1-2	Requirements finalization; telephony sandbox; slot schema approval
3-4	Call orchestrator + basic dialog manager; outbound call prototype
5-6	Integrate ASR/TTS + streaming LLM; transcript persistence; status updates
7	Pilot with 5% leads (English only); gather overrides

Week	Milestone
8	Quality tuning (prompt refinement, fallback logic)
9	Multi-language extension; add scheduling tool; dashboard KPIs
10	Rollout to 30% leads; ROI measurement; go/no-go for full deploy

14. Pseudocode Snippet (Dialog Loop)

```
while call_active:
    audio_chunk = receive_audio()
    partial_text = asr.stream(audio_chunk)
    dialog_state.update_transcript(partial_text)
    if dialog_state.needs_response():
        prompt = build_prompt(dialog_state)
        llm_resp = llm.stream(prompt)
        structured = extract_json(llm_resp)
        dialog_state.apply(structured)
        if dialog_state.complete():
            finalize_and_close()
            break
        tts_audio = tts.synthesize(llm_resp.text)
        send_audio(tts_audio)
```

15. Deployment & Scaling Considerations

- Horizontal scaling: stateless call orchestrator pods; sticky session to dialog manager instance via Redis keys.
- Peak concurrency estimation: `leads_per_day * contact_attempt_distribution`. Size ASR/LLM throughput accordingly.
- Observability: traces tagging call_id; dashboard for latency percentiles (p50, p95).
- Security: encrypt recordings at rest (AES-256), role-based access to transcripts.
- Data Privacy: configurable redaction (phone, email) at transcript layer.

16. Why It's Worth Building

Strategic advantages: faster qualification improves pipeline velocity; proprietary conversational data builds defensible ML improvements; reduces human drudgery; positions platform as intelligent acquisition system (pricing leverage). Cost profile favors automation beyond modest lead volume thresholds; extensible to cross-sell, survey, nurture sequences.

17. Summary for Leadership

An AI qualification agent can realistically cut human initial contact workload by >60%, accelerate speed-to-contact to under 5 minutes, and maintain qualification parity with calibrated scripts. Early investment (~10 weeks) leads to medium-term variable cost savings and strategic differentiation. Recommend proceeding with discovery validation (stakeholder questions) immediately and initiating a hybrid MVP build.

Document Version: 0.1 (Draft) Date: 2025-11-15 Author: AI Assistant (Based on provided dataset context)