Building Agentic Applications for the Edge

Amit Mate

Founder and CEO

GMAC Intelligence

GMAC Intelligence

Table of contents

Real-time multi-modal agentic applications

- Real-time agentic application
  - AI-driven systems that act autonomously, adapt dynamically, and make intelligent decisions in real-time
- Multi-modal
  - Agents requiring multiple modalities of data such as speech, vision, text, touch, e.g., smart-assistants
- Vertical
  - Agents specializing in narrow tasks and can operate on low-power edge platforms

*For scope of this talk, we focus on agentic applications that display the above characteristics

Low-power edge platforms for real-time multi-modal agents

- Leading low-power edge platforms for consumer-grade agentic applications, specs: {> 40 TOPS, > 8 GB DDR, < 25 W}, key metrics: { $/TOPS, mW/TOPS}
  - Jetson Orin Nano**
    - 67 TOPS (INT8), CPU: 6x Cortex®-A78, GPU: 1024-core Ampere, 8 GB DDR5 , 7-25 W
  - Snapdragon 8- Gen 3**
    - 45 TOPS (INT8), CPU: 8-core Kryo, NPU: Hexagon, GPU: Adreno , 8 GB DDR5, 7-12 W
  - Other notable vendors*
    - Apple, Mediatek (Mobile SoCs)

*Not an exhaustive list of devices or vendors

**Sampling of platforms from vendors that have demonstrated 20-40 tok/s on LLMs (8B Llama) and some VLM functionality

QSRBot: Drive-thru agent concept video

Deep-dive: Drive-thru agent (QSRBot) composition

- Context:
  - Menu, brand & location awareness, currency, language, POS systems, menu-specials, loyalty programs, personalized recommendations.
- Skills:
  - Customer Identification: Vision-based face/vehicle recognition.
  - Customer Dialog: Speech conversation (listening, talking, order-extraction).
  - Billing: Secure payments, kitchen order coordination.
  - Fulfillment & Handoff: Food preparation, packaging, robot/human-assisted handout.

Deep-dive: Drive-thru agent (QSRBot) composition

- Workflows:
  - Order Completion: menu selection, customization, billing.
  - Complaint Handling: error correction, refunds, service recovery.
  - Dynamic Promotions & Cross-Selling: consent for personalization, personalized upselling based on order history and specials, small-talk

Deep-dive: Order completion workflow

- Autonomous conversation
  - What to say?
    - Speech and language models (ASR, TTS, embedding, order-mapping)
    - Canned text and/or audio
  - When to say?
    - Visual trigger (seeing a customer/car, ANPR)
    - Audio trigger (when customer stops speaking, VAD)
    - Task trigger (when customer payment is done, payment-callback)

Deep-dive: Order completion workflow

- Dynamic adaptation
  - Menu memorization and changes
    - Vision and language models (OCR for menu text extraction, LLM for JSON conversion)
    - Json file synchronization with cloud/server
  - Prosodic TTS/rendering based on customer mood (how to say?)
    - Visual trigger (facial emotion detection)
    - Audio trigger (speech emotion detection)
    - Semantic trigger (language model-based sentiment analysis)

Deep-dive: Order completion workflow

- Real-time decision making
  - Sub-second menu extraction and bill display
    - < 500 ms reliable VAD
    - Local integration with billing system (Square, Clover)
  - Sub-second workflow adaption based on customer dialog
    - Real-time state machine (change from order-taking to complaint handling workflow seamlessly)

Deep-dive: On-device models for drive-thru agent (QSRBot)

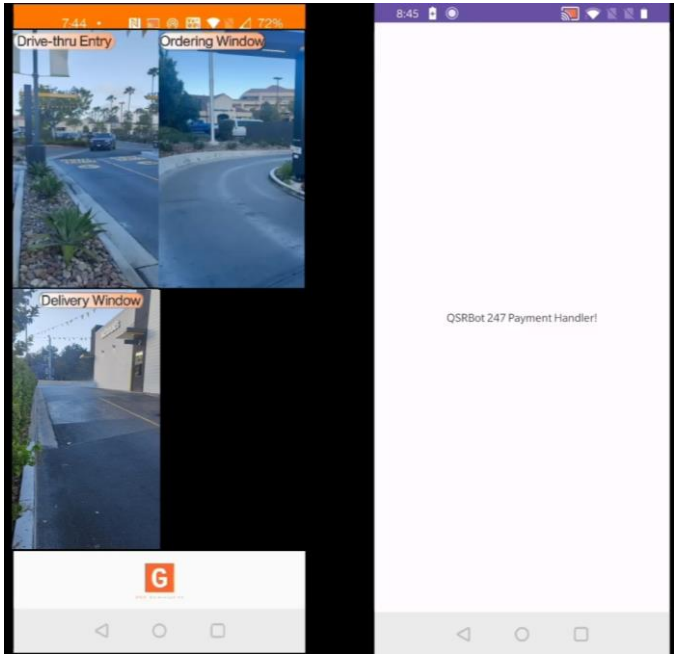
Small Models (< 100M parameters)	Modality	Platform	Workflows
SSD (MobileNetv2)	Vision	SD*, Jetson	Customer arrival trigger
OCR (MLKIT, Tao)	Vision	SD, Jetson	Customer identification/ Menu automation
ASR-English (Whisper-tiny, Moonshine, Riva-models)	Speech	SD, Jetson	Customer dialog
TTS-English (MLKIT, Riva-models)	Speech	SD, Jetson	Customer dialog
NER based classification (custom)	Language	SD, Jetson	Order extraction/Billing/Customer Dialog
Text-embedding (mini-LM-L6-v2)	Language	SD, Jetson	Customer identification

*”SD” = Qualcomm Snapdragon

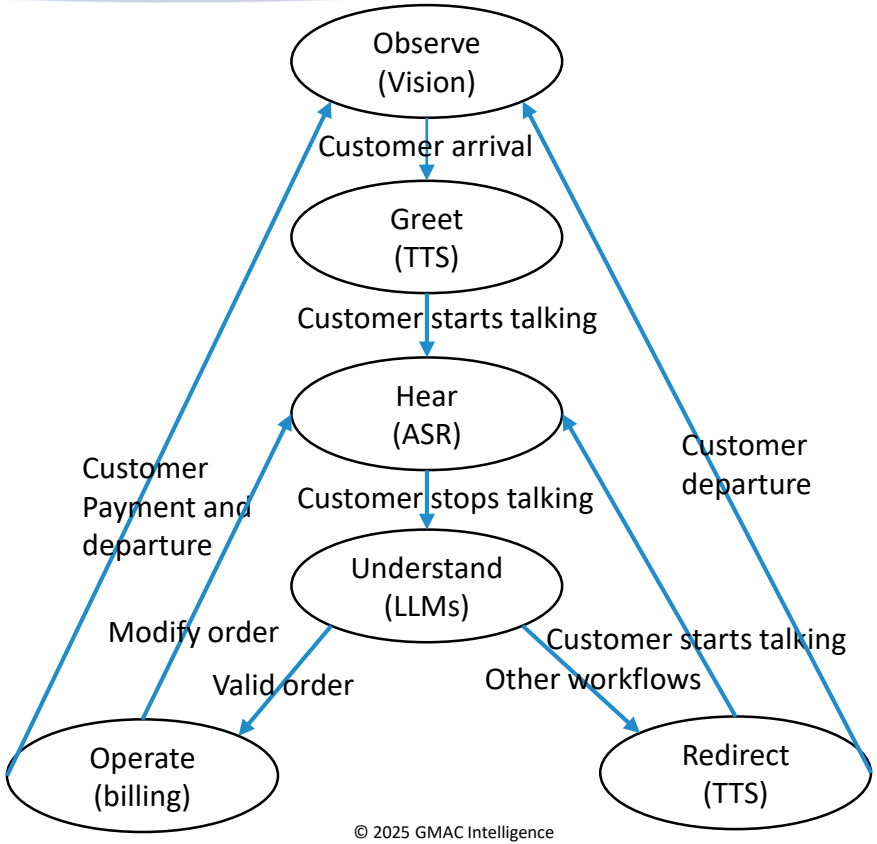
Deep-dive: On-device models for drive-thru agent (QSRBot)

Large Models (1B-10B parameters)	Modality	Platform	Workflows
Llama 3.2 2B (transformer)	Language	SD, Jetson	Customer dialog
Gemma2-2B (transformer)	Language	SD, Jetson	Customer dialog
PaliGemma2 3B	Vision+ Language	Jetson	Customer identification/ Menu automation
Gemma2 9B	Language	Jetson	Customer dialog
Gaussian splatting (NERF)	Graphics	SD, Jetson	Rendering a talking agent
Prosodic model	Vision+ Speech	SD, Jetson	Customer sentiment analysis

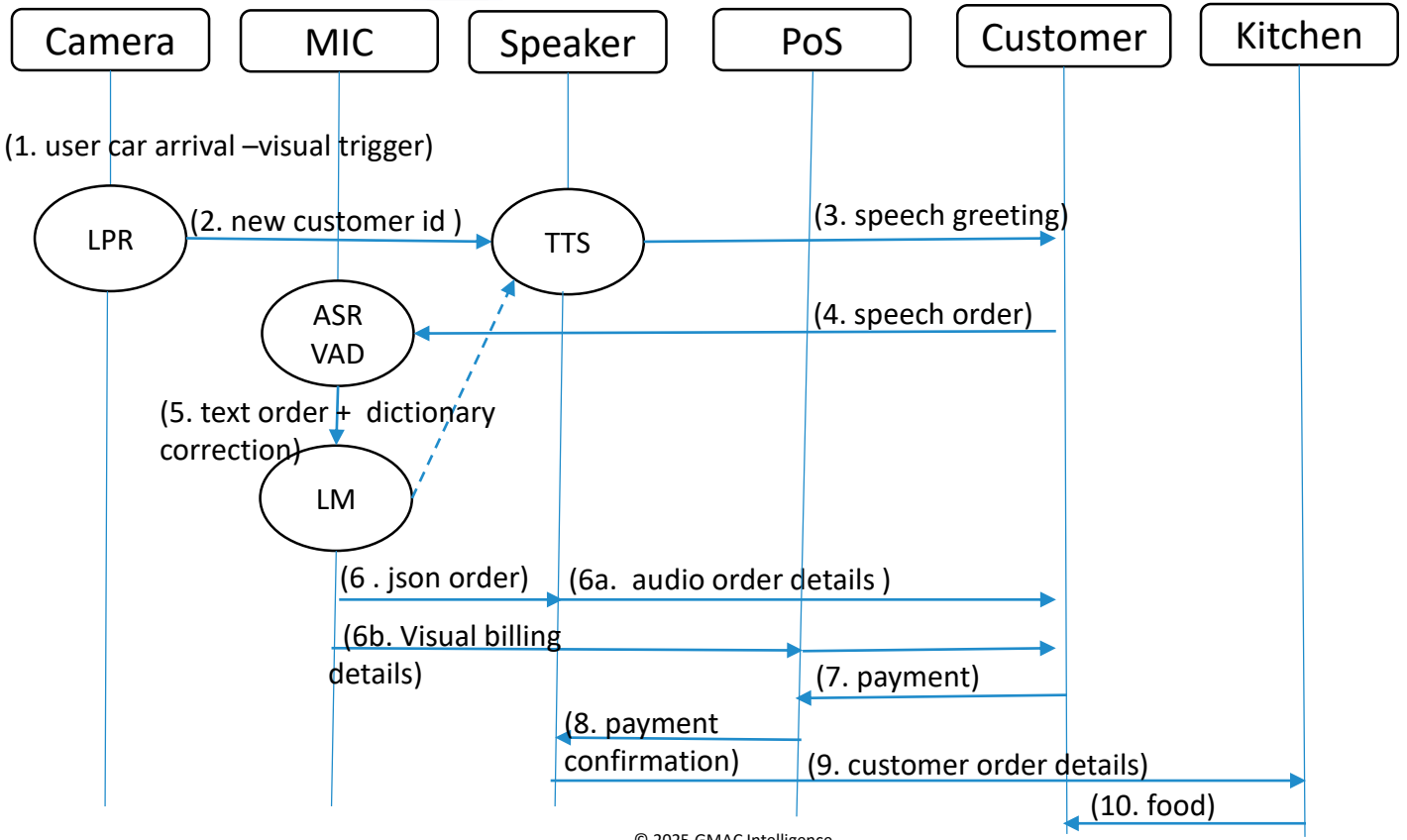
QSRBot workflow automation demo



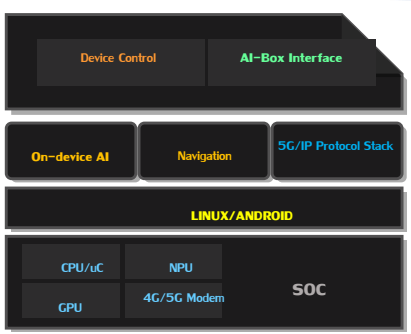
Deep-dive: QSRBot state chart



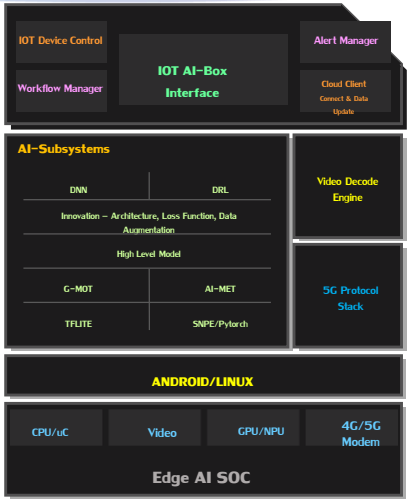
Deep-dive: Model interaction for ordering workflow



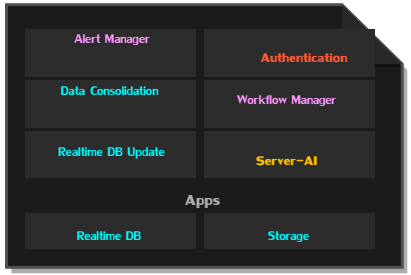
Deep-dive: QSRBot orchestration on edge



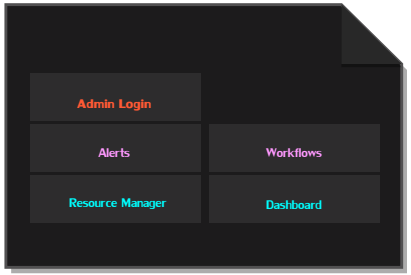
DEVICE



Edge Cloud (AIBox)



Cloud



Portal

Challenges to achieve real-time performance

- Multiple neural network run-time orchestrations (TFLITE, ONNX, PTE)
- Efficient model mapping to underlying compute (CPU, NPU, GPU, PVA)
- Hardware acceleration for pre-/post-processing of video/audio signals (cDSP, VIC, video enc/dec)
- LLM/VLM performance (accuracy/speed)
- Integrations with non-AI components is sub-optimal – lack of agentic APIs for billing etc.

Agent architecture evolution

- Composable to integrated agent architecture (more multi-modal models)
  - Current state of multi-modal: VLMs, VILA
    - Enablers : TVM, MoE architectures, distillation
  - Future:
    - Talking heads (integrated vision, speech, text input and output graphic, audio)
      - Advantages: Ease of orchestration, sophisticated, less rules
      - Challenges : Task specific multi-modal model fine tuning
    - Agentic APIs for non-AI components

Conclusion

- The bitter lesson(s)
  - “The only thing that matters in the long run is the leveraging of computation”
  - “We want AI agents that can discover/adapt like we can, not which contain what we have discovered”, i.e., going beyond the imitation game
- What do we do in the short run?
  - Use composable agents – “use what we have discovered – aka leverage rules/compute”
- What do we do in the long run?
  - Leverage edge compute with the most energy efficient multi-modal models for agents

References

Qualcomm AI-hub for models and benchmarks

https://aihub.qualcomm.com/models

NVIDIA Jetson AI-lab for models and benchmarks

https://www.jetson-ai-lab.com/

Mobile SoC AI benchmark

https://ai-benchmark.com/ranking_processors.html