SUMIT BINDRA

Surrey, BC, Canada (open to relocation) | (778) 952 5158 | sumitbindra@gmail.com
linkedin.com/in/sumitbindra | github.com/sumitbindra

Forward Deployed / Applied AI Engineer | Multi-Agent LLM & Voice AI | 15 Years Transportation & Data Modeling

SUMMARY

Transportation and data modeling expert with **15 years** of experience who **builds and ships production-grade multi-agent LLM and voice AI applications**. Since 2021 has built and launched AI products end to end: a **Gemini-orchestrated multi-agent CRM** with tool-calling agents across SMS, email, WhatsApp and voice, **DSPy-optimized prompts and an eval suite**; ElevenLabs/Twilio voice agents; document-AI extraction; an **MCP server** for travel-model data; and an **LLM-driven four-step travel demand model** (abstract submitted to MoMo 2027). Bridges deep domain knowledge (activity-based and four-step models, ridership forecasting, survey and mobility data for TransLink, state DOTs and MPOs) with hands-on AI engineering in Python, TypeScript and FastAPI, building daily with Claude Code, Codex and Cursor. Brings forward-deployed delivery habits from a multi-million-dollar TransLink program and eight years of client-facing consulting at RSG. Strongest skill is communication, sharpened by working with YC founders to turn their ideas into products.

SKILLS

AI / LLM: Multi-agent orchestration (Google ADK); tool/function calling; RAG and hybrid retrieval (embeddings + BM25); LLM evals and test harnesses; DSPy prompt optimization; MCP servers; voice AI (ElevenLabs Conversational AI, Twilio); structured extraction (Pydantic/JSON schema); prompt engineering; LLM fine-tuning (hands-on, deepening); Gemini, OpenAI (incl. Assistants API), Claude via OpenRouter; LlamaIndex

Languages & Frameworks: Python; TypeScript; React; Next.js; FastAPI; SQL; R; pytest, Vitest, Playwright

Data & Infrastructure: Supabase/Postgres (Auth, RLS multi-tenancy); DuckDB; Parquet; Docker; Railway; Modal (serverless); Databricks; pandas, GeoPandas, scikit-learn

AI Dev Tools: Claude Code; Codex; Cursor; Replit Agent

Transport Modeling: ActivitySim; AequilibraE; four-step models (gravity/IPF, MNL, Frank-Wolfe assignment); activity-based models; TransCAD 6/7; GTFS; cellphone O-D, INRIX, household travel surveys; ridership and demand forecasting

Delivery: Stakeholder discovery and scoping; translating domain problems into working software; user training and adoption; multi-million-dollar program leadership; distributed teams

Currently Learning: AWS; C++

EXPERIENCE

Founder and AI Engineer | Independent AI Application Development **Jan 2021 – Present**

ouroAI: multi-agent CRM automation platform (238 commits, 28+ PRs in two months)

- Built a platform in which a **Gemini orchestrator (Google ADK) routes work to four tool-calling sub-agents** (query, outreach, sync, reporting) that run **SMS, email, WhatsApp and AI voice** campaigns through Twilio, Resend and **ElevenLabs**, with AI replies, escalation to humans and CRM sync.
- Replaced a hand-written 3-call prompt chain with a **DSPy ChainOfThought pipeline** (channel choice, escalation, reply generation), optimized with BootstrapFewShot on 89 labelled conversations to **88.9% on a held-out validation set**, backed by an **LLM eval suite** and side-by-side comparison runner.
- Built **hybrid retrieval** (Gemini 3072-d embeddings + BM25 keyword search) over CRM contacts and per-account knowledge bases synced to voice agents; stack: **React/TypeScript, FastAPI, Supabase** (Postgres, Auth, RLS multi-tenancy), **Docker on Railway**; 50+ pytest files plus Vitest and Playwright end-to-end tests.

Other AI products built and launched

- **RealEstateISA (live at realestateisa.ai):** built and launched an AI outreach service for real-estate brokers on top of ouroAI, with a Gemini-powered sales-call grader and Follow Up Boss and GoHighLevel CRM integrations.
- **AI voice call scheduler:** inbound and outbound voice agents on ElevenLabs Conversational AI over Twilio numbers; signature-verified FastAPI webhooks handle scheduling, voicemail auto-reschedule and CRM booking via server tools; Gemini analyzes each transcript for summary, sentiment, action items and reschedule intent.
- **CICA, construction document AI:** structured extraction from bid forms, specs and proposals (OpenAI/Gemini with Pydantic/JSON-schema outputs, **LlamaIndex RAG**, Azure Document Intelligence and Reducto parsing) that flags financial, schedule, scope and legal discrepancies, with an **LLM-judged eval harness**.
- **BrickNova AI:** conversational real-estate investment assistant for Miami (Brickell) condos; a **Google ADK + Gemini 2.5 Flash orchestrator** routes to specialist agents that use **tool calling** for unit search, building comparison, ROI and cost-to-own; Next.js 15/TypeScript, FastAPI, Supabase, Railway.

- **Answr conversational surveys and interviews:** AI survey and interview apps (employee satisfaction, customer and transit-rider feedback) on the **OpenAI Assistants API** and LlamaIndex workflows, with LLM follow-up probing, analysis dashboards and WebRTC voice versions; also a **multilingual (English/Spanish) HR project management app** built within weeks of the Assistants API's November 2023 launch.
- Partnered with YC founders to turn early ideas into working products, translating fuzzy requirements into clear specs and shipped features.
- Build daily with AI coding agents (**Claude Code**, with co-authored commits, CLAUDE.md and custom skills; **Codex; Cursor; Replit Agent**) to ship tested, containerized services quickly.

Project Manager and Senior Modeler | TransLink, British Columbia, Canada **Sep 2019 – Present**

- Co-developed TransLink's regional **trip- and activity-based travel demand model in Python**; key role in **ridership forecasts** for the **Surrey-Langley SkyTrain** and station- and line-level forecasts for the entire Metro Vancouver transit system.
- Lead a **multi-million-dollar household trip diary and screenline data program** covering sensitive personal information from **65% of the region's households**, owning the flow from collection through model calibration and estimation; built the supporting **data pipelines in Python and Databricks**.
- **Implementing LLMs** and other innovative solutions in modeling and customer service.
- Bridge technical teams and stakeholders: run recurring model user group meetings, build consensus on methods, and turn forecasts into clear recommendations for decision-makers.

Consultant and Data Analyst | Resource Systems Group Inc. (RSG), White River Junction, VT **Aug 2011 – Aug 2019**

- Delivered statewide and regional models for **Michigan, Tennessee and North Carolina DOTs and MPOs (SMTC, GTC)**, working directly with client agency staff from scoping through implementation, training and support.
- Built **machine learning trip generation** from a large multi-agency travel survey (bootstrapping, probabilistic sampling, GLMs) and **intermodal freight** components (rail, truck, port) in **Python and R**.
- **Cut statewide model run time 30%** through parallelization and more efficient components; added toll modeling, Value of Time sensitivity and 13 vehicle classes; analyzed **100+ projects** for multi-million-dollar prioritization.
- Analyzed a **large-scale cellphone origin-destination dataset** (TRB 2016); built GTFS transit networks, INRIX speed validation and post-processors producing agency-ready maps and analyses.

AI X TRANSPORTATION PROJECTS

LLM-Driven Four-Step Travel Demand Model (built on TypeSafe's LLM decision engine)

- Built a four-step model in which **trip generation, distribution and mode choice are answered by an LLM decision engine** instead of estimated coefficients; validated on the **Sioux Falls** region against ACS, LODES, NHTS and traffic counts and compared with a conventional gravity/MNL model (67 coefficients) under transit, parking and capacity policy tests. **Abstract submitted to Modeling Mobility (MoMo) 2027.**
- Engineered reproducible LLM experiments: hash-keyed SQLite request cache, cost guard with dry-run, placebo-controlled prompt-framing tests, temperature scaling and a calibration ladder (171 commits, 26 pytest files).

tdm-platform: Closed-Loop Travel Demand Model on Serverless Compute

- Region-agnostic **ActivitySim + AequilibraE** platform (Frank-Wolfe user-equilibrium assignment with skim feedback) that builds a model region from open OSM/TIGER data (demo: 168 TAZs, 73k links) and runs on **Modal for about $0.07 per scenario**.

travel-mcp-server: LLM Access to Travel-Model Outputs

- **Model Context Protocol server with 8 tools** that lets LLM clients query travel-model outputs (Parquet tables, OMX skims) with **DuckDB SQL**, pull trip statistics and skims, and generate charts; packaged in Docker.

EDUCATION

Master of Science | University of Connecticut, Connecticut, USA 2006 – 2007

Bachelor of Technology | Indian Institute of Technology Guwahati, India 2001 – 2005

PUBLICATIONS

- Using Cellphone O-D Data for Regional Travel Model Validation. Annual Meeting of the Transportation Research Board, Washington, D.C. (2016).
- Predicting Segment-Intersection Crashes with Land Development Data. Transportation Research Record: Journal of the Transportation Research Board (2009). Co-authors: J. N. Ivan and Thomas Jonsson.
- Explaining Road Speeds with Spatial Lag and Spatial Error Regression Models. 86th Annual Meeting, TRB (2007). Co-authors: J. K. Hackney, M. Bernard, and K. W. Axhausen.