

MAOR BRIL —

PRINCIPAL ENGINEER · AGENTIC AI

San Jose, CA | maor.bril@gmail.com | maor-bril.xyz | github.com/MaorBril | linkedin.com/in/maorbril

SUMMARY

Twenty years building production infrastructure at Google, Netflix, and Coinbase; the last two leading technical and product strategy for agentic AI at Character.ai. I build the agent systems, the Go SDK five production systems run on, and the self-improving eval wall that gates them — at **20M+ MAU**. Four of those systems are shipped interactive entertainment, including an agent-run detective game and a live murder-mystery game I built and operate alone.

EXPERIENCE

Principal Software Engineer · Character.ai

Apr 2024 — Present

- Own technical and product strategy for agentic AI across generative entertainment products reaching **20M+ monthly users**; set the org-wide standard for how agentic systems are built, evaluated, and continuously improved.
- Built **AgentX** and **AgentZero**, a fully agentic multimodal pipelines: a planner agent decomposes the request, executor agents generate image, video and audio, and reviewer agents gate every stage on trained eval scores. **30+ agentic turns, multiple specialized tools, no human in the loop** after the prompt.
- Built **JudgeJudy**, the multimodal eval wall — AI judges with custom rubrics, trained scoring models, automated metrics across text/image/audio/video, and a human calibration loop that measures AI-human score divergence and rewrites rubrics automatically. And then distilled it into a single video judge that is a fine-tuned Qwen3-VL trained on pairwise preference with Bradley-Terry loss, scoring a clip in **~3 seconds** so it can live in CI. Every release is gated on regression.
- Built **Judge Dredd**, the self-improving layer above it: an MCP-backed harness that pulls Amplitude and Statsig telemetry, correlates offline eval scores against real user behavior, identifies which quality dimensions actually predict engagement, and feeds validated gates back into the runtime — no manual tuning.
- Designed and open-sourced **claude-agent-sdk-go**, the production Go SDK underneath all of it: streaming, tool registries, hooks, MCP support, bounded goroutine pools, context propagation. **Five production systems** run on it.
- Set AI development standards across the engineering org — eval frameworks, safety guidelines, sandbox isolation, toolchain practice — enforced at deployment.
- Infrastructure: Spanner, ClickHouse, AlloyDB and WarpStream processing **5 GB/sec** of multimodal content in real time.

SHIPPED INTERACTIVE · Case Files — an agent-run detective game (Character.ai Interactive Originals). Agents generate a coherent mystery, hold character through free-text interrogation, keep world state consistent across evidence and testimony, and stay solvable — under a timer, with one accusation.

StoryLab — the agentic harness behind c.ai/fm: premise to a published, fully-voiced 20-minute episode, with a reviewer agent checking continuity against a series bible across ten-episode LitRPG arcs.

c.ai/studio — shot-level production floor for vertical microdrama: agents beat out the episode, generate multiple takes per shot, and assemble a cut a human can re-order and re-submit. Weekly release cadence.

Jeeves — agentic onboarding assistant adapting to user behavior on the same SDK; different goal, different tool registry. The platform is general, not one-off.

Senior Staff Software Engineer · Coinbase

2022 — 2024

- Led company-wide technical strategy for the Datastores org — RDS, Aurora, MongoDB, ClickHouse, DynamoDB, Redis, CockroachDB. Built and ran the Datastores and Observability teams, growing them to **20+ engineers**.
- Replaced an external observability vendor with an in-house system at **90% cost savings**.
- Architected Wallet-as-a-Service infrastructure; launched ClickHouse and CockroachDB as company-wide paved paths. Systems process **billions in daily trading volume**.

Senior Software Engineer · Netflix

2021 — 2022

- Self-Service Datastores Platform and the Data Abstractions team, saving **10–20 dev hours per week** across the org and serving **200M+ members**.

- Contributed to Netflix OSS (Hollow — delta-based in-memory sync).

Senior Software Engineer · Google Cloud

2019 — 2021

- Shipped a Cloud Spanner optimization cutting latency **up to 80%**. Rebuilt GCP Networking Quota Management; systems processing **millions of QPS** globally.

Staff / Tech Lead Software Engineer · VMware (acq. Arkin)

2015 — 2019

- Tech-led a 6-person team from zero to GA in 6 months: datastores, stream processing, compute, identity, and air-gapped enterprise readiness.

BUILT AND OPERATED SOLO

AI She Wrote · aishewrote.com

Live in production

- A live agentic murder-mystery game — world design, agent architecture, front end, and operations, all mine. Every suspect is an agent with private knowledge, a motive to lie, and a story that has to hold up under free-text interrogation while the mystery stays solvable no matter what the player types.

OPEN SOURCE

clauder

github.com/MaorBril/clauder · clauder-ai.dev · Go · MIT · 52★

MCP server giving AI coding agents persistent memory and real-time inter-agent communication across sessions and repos. Solves context loss in long-horizon agentic workflows.

agentic

github.com/MaorBril/agentic · Go · MIT

Runs Claude Code on any model, with a budget: a local multi-provider router with an LLM-classifier tier router, per-token metering and pricing, and hard-stop budgets. Static binary, no daemon.

JudgeJudy

github.com/character-ai/judgejudy · Go · MIT

Multimodal eval harness: trained scoring models, AI judges, automated metrics across text/image/audio/video, human scoring, and a self-improving calibration loop.

claude-agent-sdk-go

github.com/character-ai/claude-agent-sdk-go · Go · MIT

Production Go SDK for agentic applications: streaming, tool registries, MCP, bounded concurrency. The platform primitive behind multi-agent orchestration at Character.ai.

TALKS & WRITING

- **Why Your Video Judge Rewards Slop — and How to Fix It.** Conference talk on pairwise preference, Bradley-Terry training, and running a video judge as a CI regression gate.
- **Harnessing Data at Scale: Character.ai's Transition to WarpStream.** Podcast interview by Confluent on 5 GB/sec multimodal ingest.

TECHNOLOGIES

AGENTIC AI

Multi-agent orchestration, planner/executor/reviewer agents, agent runtimes, MCP, tool integration, self-improving eval systems, trained scoring models, preference fine-tuning (Bradley-Terry), feedback loops, sandbox isolation

POST TRAINING

Post trained (SFT and RL) models for specific agentic use cases, in order to serve these experiences at scale and quality

LANGUAGES

Go, Python, Java, TypeScript and English (What Coding Agents speak)

AI TOOLING

Claude Code, Cursor, Codex, AgentSDK, Anything else

INFRASTRUCTURE

Kubernetes, Spanner, ClickHouse, WarpStream, AlloyDB, DynamoDB, Redis, CockroachDB, MongoDB, Kafka, Terraform

EDUCATION

BSc, Electrical Engineering & Communications

2008 — 2012

Holon Institute of Technology (HIT), Israel — graduated while serving as an officer in the IDF Intelligence Corps.