

ARDA BASARICI

AI Engineer — LLM systems, evaluation-first

Software engineer, 3 yrs production · Mathematics · Turkey (UTC+3) · open to remote (contract / EOR) and relocation

ardabasarici@gmail.com · ardabasarici.dev · linkedin.com/in/ardabasarici · github.com/arda-basarici

SUMMARY

Software engineer with 3 years shipping production games and a mathematics background, now focused on evaluation-first LLM systems: deterministic where a model is not needed, measured against ground truth where one is, with the residual error disclosed. Built and shipped SteamLens, a live LLM review-analysis app, from a human-adjudicated gold set to approval-gated deployment, now on infrastructure held in code. Claude Code daily on every project; each change accepted only after tests, CI and expected behaviour pass.

PROJECTS

SteamLens — evaluation-first LLM review-analysis app

Jul – Aug 2026

Python · FastAPI · SQLite · multi-provider LLM APIs (DeepSeek, Gemini) · Docker · GitHub Actions · Cloudflare

steamlens.ardabasarici.dev · github.com/arda-basarici/steam-lens

- **LLM evaluation:** measured the production labeller against a 250-review gold set (assist-model drafts, every label adjudicated by a single annotator before any candidate model was scored, the assist model excluded from the candidates): **F1 0.766 (95% confidence interval 0.713–0.811)**; a cross-family LLM judge, calibrated against the same gold set under a predefined rule, extends the check beyond it; rejected a higher-scoring candidate model at ~12× the cost.
- **Grounding / hallucination control:** quotes verbatim-checked at write time (~2.9% of attempted quotes rejected; **0 non-verbatim among 163,842 stored evidence spans**), quotes and numerals re-verified at compose time; the grounding gate blocks any unverifiable numeral or quote before publish; residual attribution error measured (11.6%) and disclosed in the product.
- **Production system and operations:** built and shipped, solo, the end-to-end pipeline from review sampling and structured LLM classification through deterministic aggregation to grounded report generation (every displayed number computed from the labels, never by the model): a 135,260-review census labelled for \$3.80 (one refusal logged); sampling validated against census ground truth on a held-out game; CI evaluation gates, approval-gated deployment with rollback, per-report spend admission, public ops dashboard.

Platform — the operations layer under the deployed projects

Aug 2026 – Present

Terraform · Ansible · AWS (EC2, SSM, IAM OIDC, CloudWatch) · Cloudflare · Caddy · Docker · SOPS/age

github.com/arda-basarici/platform

- **Infrastructure as code, adopted on zero-diff plans:** extracted the shared operations layer under two deployed projects (Cloudflare edge, a self-managed VPS, an AWS host) into its own repository: the DNS records, the deliberately set edge configuration and the AWS stack in Terraform (hand-made resources adopted on zero-diff plans); the VPS is rebuildable from a blank host by an Ansible play with an acceptance script.
- **Recovery proven, not asserted:** a rebuild-and-restore drill from a blank host and a blank control node to SteamLens serving its restored production backup on a fresh host, in about an hour; a monthly automated restore check; the alarm path verified by forcing the alarm, which found and fixed a silent SNS policy refusal.

PROFESSIONAL EXPERIENCE

Tiplay Studio — Senior Game Developer

Feb 2022 – Dec 2024

Remote · joined as Game Developer, promoted to Senior · ~5-person web-games team

- **Delivery and cross-project leadership:** shipped 20+ production games for web, in-car and TV platforms (AirConsole, CrazyGames, Famobi, Snapchat, BMW / VW) using JS/TS, C#, Unity WebGL, Cocos Creator and PlayCanvas; reviewed and pair-programmed across the team's projects and developed platform approaches and workarounds reused by other developers; promoted to Senior for cross-project problem ownership.
- **Product ownership, live multiplayer (TypeScript, Cocos Creator):** developed Card Lounge largely solo, a live multiplayer card-game hub (poker, blackjack, gin rummy, hearts) for AirConsole's in-car platform: phone-controller multiplayer, bots for single-player, leaderboards, daily missions; over a year in production.
- **Real-time backend and production support (WebSocket):** designed and shipped AirConsole's first car-to-car multiplayer system, including the server and WebSocket backend the platform did not provide, taking it from experiment to production; the platform also routed other studios' integration problems to Tiplay, including a live production issue resolved on a previously untested BMW model at Gamescom 2024.
- **Platform breadth (PlayCanvas, Unity WebGL):** an asynchronous social multiplayer game for Snapchat in PlayCanvas; mobile-to-web ports for CrazyGames in Unity WebGL, adapting sizing, levels and monetization to the platform and optimizing browser performance.

Independent development; compulsory military service

2025

- Led weekly problem-solving sessions and office hours across 5 semesters and 7 undergraduate courses (calculus, advanced calculus, linear algebra, mathematics for social sciences); graded exams and written solutions for rigour and clarity.

ADDITIONAL PROJECTS

Blackjack RL — RL agents audited against a provable optimum

Jun – Jul 2026

Python · PyTorch · NumPy · 3 reports on ardabasarici.dev · github.com/arda-basarici/blackjack-rl

- **Policy audit against a provable optimum (RL):** audited a tabular Monte-Carlo agent against provably optimal basic strategy (~93% of state-action cells rediscovered; residual traced to rare-state coverage and near-ties), then a DQN on the same task (~82% agreement for the naive network, ~91% with a stabilizer stack, still below the table; the cost located at the policy's sharp boundaries by an input-encoding ablation), on a from-scratch simulator validated against the published 0.45% house edge before use.
- **Experiment design / falsification:** a DQN bet-sizer never learned Kelly betting on a measured counting edge; the same network learned it on a denoised reward, and the "keyed on wealth, not edge" hypothesis was falsified by two controlled experiments; the negative result is the published finding.
- **Reproducibility / provenance:** every run persists seed, config and git hash; the betting report's headline numbers asserted against run data at build time; Wilson-score intervals for ruin and drawdown.

Learned A* heuristic — ML heuristics vs Manhattan distance

Jun – Jul 2026

Python · scikit-learn (gradient boosting) · NumPy

ardabasarici.dev/reports/pathfinding_report.pdf · github.com/arda-basarici/pathfinding-ml

- **Supervised ML, leakage-safe evaluation:** a learned cost-to-go heuristic (gradient boosting) on a whole-maze train/test split: at the report's 50/50 maze mix, 17.3% fewer node expansions (not wall time) at a 0.2% mean optimality gap on held-out mazes.
- **Confound detection, distribution-shift testing:** the pooled result looked like a wash; splitting by maze type exposed opposite effects, training on pure maze-style subsets traced the effect to training-distribution composition rather than the added feature, and a cross-style transfer test exposed the failure modes under shift; the first explanation was revised on the evidence.

Steam Reviews — resumable pipeline + within-game statistical analysis

Jun – Jul 2026

Python · pandas · Parquet · pandera · scipy

ardabasarici.dev/reports/steam_review_report.pdf · github.com/arda-basarici/steam-reviews

- **Resumable data pipeline, schema contracts:** ingested and validated 298,553 reviews (50 games) with at-least-once fetch semantics and pandera contracts gating Parquet promotion; no text modelling.
- **Simpson's paradox guard, confirmatory modelling:** required each of the four reported findings to reproduce game by game; one 42-point pooled effect failed and is reported as a discard.

SKILLS

AI / ML	LLM evaluation (human gold sets, LLM-as-judge calibration, bootstrap and Wilson CIs, eval gates in CI), prompt and structured-output design, grounding and hallucination control, cost and latency control, PyTorch, scikit-learn, reinforcement learning, experiment design
Backend / data	Python, FastAPI, SQL (SQLite), pandas, NumPy, Parquet, REST APIs, SSE, WebSockets
Infrastructure	AWS (EC2, SSM, IAM), Terraform, Ansible, Docker, GitHub Actions CI/CD, Cloudflare, Caddy, SOPS/age, Linux
Mathematics	probability and statistics, linear algebra, combinatorics, mathematical modelling
Programming	Python, TypeScript / JavaScript, C#
Spoken languages	Turkish: native · English: full professional

EDUCATION

Boğaziçi University — B.Sc. Mathematics

2014 – 2018

GPA 3.40 / 4.00, Honors Certificate; third-highest GPA in the Mathematics cohort · TÜBİTAK undergraduate fellowship

Boğaziçi University — M.Sc. Mathematics coursework

2018 – 2022

GPA 3.92 / 4.00; coursework complete, thesis not submitted · TÜBİTAK graduate fellowship 2018 – 2020