

SALAH-EDDINE BENBRAHIM, PhD, MBA, B.Eng.

Curriculum vitae · AI evaluation, trustworthiness and reproducibility · September 2026

• benbrahim.salah.eddine@cineinfinite.com • +1 647-875-0826 (CA) • +33 7 81 64 82 93 (FR) • [LinkedIn](#) •

cineinfinite.com • [CineInfinite Studio \(live\)](#) • [GitHub](#) • [Hugging Face](#) • [ORCID 0009-0005-7050-4014](#) • [DBLP](#) • [PolyPublie](#)

RESEARCH STATEMENT

My research asks a single question: how do we know that a number used to judge an AI system can be trusted? I work on the reliability, reproducibility and comparability of AI evaluation — the measurement layer beneath every capability and safety claim. The premise is that a safety property which cannot be measured reliably cannot be assured, and that the judgements on which such measurements rest — human or model-produced — are routinely assumed stable rather than shown to be.

Two sole-authored papers, currently under review, establish that this assumption fails in measurable and detectable ways. Alongside them I build and operate the instrument: CineInfinite, a live multimodal media governance platform that produces reproducible audit evidence — hashed inputs, seeded runs, bootstrap intervals, multiplicity correction, calibration error, and a mandatory limitations section, under the rule that every value traces to an artifact and missing data is stated rather than invented.

My work draws on two careers. I completed a doctorate in mathematical optimization, spent a decade shipping production systems for institutions that could not fail quietly, and now bring both to research. That combination is the source of the work: formulating a mixed-integer program teaches that an objective and its constraints must be written down exactly, or a solver returns a confident wrong answer. A published reliability coefficient is the solution to a problem whose constraints were never stated — and two such figures are routinely compared as though they had solved the same problem.

RESEARCH INTERESTS

Trustworthy and responsible AI · evaluation, benchmarking and measurement validity of AI systems · reproducibility and metrology of machine learning · reliability of human annotation and preference data · LLM-as-a-judge and the evaluation of AI evaluators · AI audit, assurance and conformity assessment (EU AI Act, NIST AI RMF, ISO/IEC 42001) · multimodal and AI-generated content assessment · mathematical optimization for resource allocation under constraints · distributed systems and cloud infrastructure.

EDUCATION

Doctorate (PhD), Computer Engineering 2016

Polytechnique Montréal, Université de Montréal

Master of Business Administration (MBA) —

Université du Québec à Montréal (UQAM)

Master of Applied Science (M.Sc.A.), Computer Engineering 2011

Polytechnique Montréal

Thesis on defence against traffic-analysis attacks in wireless sensor networks.

Bachelor of Engineering (B.Eng.), Computer Engineering —

Polytechnique Montréal

PUBLICATIONS

Under review — sole author

1. **Index terms: inter-rater reliability; variance components and identifiability; human reference panels; LLM-as-a-judge evaluation; annotation budget allocation; calibrated diagnostics; reporting protocols.** Under review, 2026.
2. **Index terms: inter-rater reliability; construction choices in reliability estimation; rater screening and aggregation; panel size; multiverse analysis; reproducibility of published coefficients. Reproduction package: doi.org/10.5281/zenodo.22698801.** Under review, 2026.

Peer-reviewed journal articles

3. Benbrahim, S.-E., et al. "Live Placement of Interdependent Virtual Machines to Optimize Cloud Service Profits and Penalties on SLAs." *IEEE Transactions on Cloud Computing*, 2019. **15 citations.** Mixed-integer programming and Tabu Search for SLA-constrained placement and migration; validated on Google traces; 13.5% avg. reduction in service penalties and up to 37% reduction in physical-machine usage.
4. Perrier, N., Benbrahim, S.-E., et al. "A comparison of project control standards based on network analysis." *International Journal of Information Systems and Project Management*, 2019. 11 citations.

Peer-reviewed conference papers

5. Perrier, N., Benbrahim, S.-E., et al. "The core processes of project control: A network analysis." *Procedia Computer Science* (CENTERIS 2018). **6 citations.**
6. Benbrahim, S.-E., et al. *Network and Information Systems (NBIS)*, 2014. **3 citations.**
7. Benbrahim, S.-E., et al. *Asia Modelling Symposium (AMS)*, 2014 — two papers.

Theses

8. Doctoral dissertation on real-time migration of interdependent virtual machines in cloud networks, Polytechnique Montréal, 2016.
9. Master's thesis on defence against traffic-analysis attacks in wireless sensor networks, Polytechnique Montréal, 2011. **4 citations.**

Bibliometrics

Nine records · **41+ citations** · **h-index 4**. Most cited: *IEEE Transactions on Cloud Computing* (15) · *IJISPM* (11) · *Procedia Computer Science* (6). Full record: PolyPublie, DBLP, ORCID.

RESEARCH SOFTWARE AND INFRASTRUCTURE

CineInfinite — multimodal media governance platform 2023 – present

Founder and architect · designed and built with AI-assisted development · live at cineinfinite.com

- Governs **14 evaluation corpora** and **1,435,331 rated assets** across video, image, speech and text, reconciling seven distinct rating conventions (Likert 1–5, Likert 1–4 over five dimensions, DMOS 0–100, MOS, MOS 0–100, empirical min-max, WMT direct assessment) under an explicit scale contract that flags non-standard scales rather than normalizing them silently.
- **24 scoring modules across 6 categories, 134 features**: quality (16 modules — CLIP text-video alignment, optical-flow motion coherence, interpolation-artifact detection, subject and identity consistency, world-model surprise, causal reasoning, aesthetic composition, long-term narrative coherence), published baselines (DOVER, ICCV 2023; FAST-VQA, ECCV 2022), multimodal safety (CLIP zero-shot NSFW, violence, distress, weapon and AI-artifact detection), AIGC origin detection (RandomForest over 16 signals, plus a DINOv2 cross-check), Shapley-like explainability, and benchmark fusion.
- **Statistical methodology**: SRCC as primary metric with PLCC after VQEG logistic fit, KRCC and RMSE; 95% bootstrap confidence intervals over 1,000 resamples; Benjamini–Hochberg FDR correction across features; calibration reliability over ten prediction-quantile bins with mean absolute calibration error reported *including when it is poor*.
- **Cross-dataset generalization analysis** over 134 features × 12 datasets, surfacing **99 sign-inversion integrity flags** — features whose correlation with human judgement reverses between corpora.
- **LLM-as-a-judge module** measuring judge–human agreement across providers, with an inter-judge leaderboard reporting SRCC, PLCC and MAE against human MOS.
- **Reviewer-safe report generation** — reproducibility appendix, draft paper and architecture paper, each exportable as Markdown, LaTeX, DOCX and PDF, assembled from real run artifacts under an enforced rule: every value traces to a manifest, metrics file or audit, and missing data is stated, never invented. Each appendix carries a mandatory limitations section.
- **Verifiability and replayability**: MOS-file SHA-256 with a verification endpoint; video-set fingerprinting over a content manifest; and per-run metadata comprising git commit, git-diff size, seed, configuration hash, calibration hash, requirements hash, MOS hash and execution plan.
- **AI-assisted, human-governed**: methodology, architecture and statistical framework designed by me; the platform is implemented with AI-assisted development under a verification protocol — the same discipline it applies to the systems it evaluates.
- **Engineering discipline**: over 90% test coverage executing in under 30 seconds; under 5% measurement variance across repeated runs; computation-level audit logging; semantic versioning with full backward compatibility; automated CI/CD with go/no-go release gating; PyPI packaging; Zenodo DOI archival.
- **Attribution**: every dataset and model cited to its original authors, venue and licence — 12 datasets and 6 models across QoMEX, CVPR, ECCV, ICCV, EMNLP, ACM MM, IEEE TIP, IEEE Access, Odyssey and WMT, with licences tracked individually.
- **Accessibility**: 17 operational surfaces; 18 interface themes; five languages including Arabic right-to-left.

Adjacent artifact

- **The Left Hand — instruction-driven cultural adaptation**: one poem, ten versions, six languages (English, French, German, Spanish, Hebrew, Arabic), with nine invariants specified in advance and each version's adaptation decisions documented. Built as a controlled artifact for studying how generated text is judged — the production side of the evaluation work above.

RESEARCH AND ACADEMIC EXPERIENCE

Postdoctoral Research Fellow — Polytechnique Montréal 2017 – 2018

- Network analysis of project-control processes in information systems; published at CENTERIS 2018 and in the *International Journal of Information Systems and Project Management*.

Doctoral Researcher and Lecturer — Polytechnique Montréal

2009 – 2016

Department of Computer Engineering and Software Engineering · LARIM

- Doctoral research on mixed-integer programming and Tabu Search metaheuristics for SLA-constrained cloud resource allocation, validated on simulations of 25,000+ VMs.
- Taught MATLAB programming to first-year engineering undergraduates.

PROFESSIONAL EXPERIENCE

Technical Lead and Enterprise Architect — Government of Quebec

2023 – 2025

Public digital platform serving more than two million citizens

- End-to-end architecture of government digital platforms: 40% latency reduction (2.5s to 1.5s), 99.9% availability, eight major releases delivered on schedule with zero critical incidents. Microservices, REST, cloud infrastructure, CI/CD.

Founder and AI Architect — JokeJoy2023 – 2025

Multi-LLM production SaaS

- Multi-LLM system in production with automatic provider routing on latency and cost thresholds (GPT-4o primary, Gemini 1.5 fallback).

Senior Business Intelligence Analyst — Government of Quebec

2019 – 2020

- SAP BI, data governance and data quality; 35% reduction in data-quality issues; mentored a team of eight analysts.

TEACHING

- **MATLAB programming for first-year engineering undergraduates** at Polytechnique Montréal.
- Languages of instruction: French and English.

SERVICE TO THE PROFESSION

- Open-science practice: all research software released open-source with DOI archival and licence-tracked attribution of every third-party dataset and model.
- Available for peer review in AI evaluation, reproducibility, multimedia quality assessment and cloud optimization.

TECHNICAL COMPETENCIES

Evaluation & statistics : measurement validity · inter-rater reliability (Krippendorff, Fleiss, ICC) · bootstrap confidence intervals · multiplicity correction (Benjamini–Hochberg) · calibration analysis · multiverse and specification-curve analysis · SRCC / PLCC / KRCC

Machine learning : PyTorch · ONNX · CLIP · DINOv2 · ArcFace · LLM evaluation and LLM-as-a-judge · multi-provider orchestration · RAG

Optimization & OR : mixed-integer programming · Tabu Search and metaheuristics · constrained resource allocation · SLA modelling

Engineering : Python (async, concurrent) · pytest · GitHub Actions · semantic versioning · PyPI · Docker · distributed systems · REST and microservices · AWS · GCP

Governance & compliance : EU AI Act conformity assessment · NIST AI RMF · ISO/IEC 42001 and 42006 · audit evidence and reproducibility appendices · data provenance and integrity verification

LANGUAGES

French — native · **English** — professional (fluent). Able to write, teach and present in French and English.

WORK ELIGIBILITY

I can work in Canada, France and across the European Union and EEA. Eligible for TN status under the USMCA for the United States. Available in Montréal, Toronto and Paris.