BAD | BARCELONA AI & DATA SYSTEMS

BAD — Barcelona AI & Data Systems

BAD Meetup #2: AI Born in Barcelona

Practical applications, technical deep-dives, innovative research happening all around Barcelona, and an honest discussion on how Barcelona can become Europe's AI and Data hub.

15th Sep. 2026, Glovo Yellow Park

Hi, It’s Ema! Welcome to Glovo 👋

Emanuele Bardelli

Data Director · Glovo

- Leads the Data Foundations team at Glovo, empowering foundational Data Domains via advanced analytics, and enabling Data & AI capabilities to augment business productivity through automation

And welcome to the BAD Community 🚀

Barcelona AI & Data Systems

Building the community where Barcelona's academia, startups, and enterprises shape the cloud and data future for AI together.

www.barcelonadata.com

BAD | BARCELONA AI & DATA SYSTEMS

Join the conversation on the BAD Discord 💌

- Join the BAD Community Discord server, read and accept the #rules to become a member
- Share your thoughts and questions about today in #meetup-02-sep-26
- Keep the conversation going and help us build the community in #all-baddies
- Join the next meetup in November at UPC #meetup-03-nov-26

BAD | BARCELONA AI & DATA SYSTEMS

You’re in for a treat 🍭

BAD Applications: Keyless access to LLM Gateway — Jean, Carlos @ Glovo

BAD Research: Human Gaze as an alignment signal for LLMs and VLLMs — Ángela @ Telefónica

BAD Systems: Towards an AI Software Factory for Data Systems — Laura, Subru @ Microsoft

BAD Panel: AI built in Barcelona, for the World — Emanuele, Laura, Carlos, Jesus

TALK

BAD | BARCELONA AI & DATA SYSTEMS

Keyless access to LLM Gateway

No keys, no drama

Carlos Casanova (Security Engineer), **Jean Baudin** (Staff Software Engineer) · Glovo

Jean Baudin

Staff Software Engineer

AI and Kubernetes Expert at Glovo

Carlos Casanova

Security Engineer II

Securing AI implementations across all Glovo's SaaS products

What is an AI Gateway?

Multiple vendors through a single endpoint

BAD | BARCELONA AI & DATA SYSTEMS

AI Gateway is the Single Control Plane for LLMs

Reverse Proxy

Specialized reverse proxy optimized specifically for orchestrating and managing language models.

Unified API

A single, OpenAI-compatible interface bridging access to 100+ LLMs, including OpenAI, Anthropic, AWS Bedrock, and Vertex AI.

Landscape

Leverage open-source and enterprise-grade frameworks: **Kong · Cloudflare · Bifrost · LiteLLM**.

BAD | BARCELONA AI & DATA SYSTEMS

Why Do We Need It?

- **Cost Controls** — Team, automation, and client-level budgets
- **Guardrails** — Active defense against PII leaks, jailbreaks & prompt injection
- **Policy Engine** — Enforceable rulesets triggered in real time
- **Usage Telemetry** — Granular insight into employee vs. customer-facing AI
- **Model Governance** — Seamless switching between closed and open-weights models

BAD | BARCELONA AI & DATA SYSTEMS

Why LiteLLM? Architecture Decision

Time to Value — Rapid deployment in days, not quarters

Open Source — Full transparency, self-hostable, zero vendor lock-in

Product Fit — Lightweight and purpose-built for multi-provider routing

The Long-Term Challenge: Token Management

Everything's fine, right?

The Hidden Pitfalls

User-to-Machine (U2M) — Unmonitored token generation

Machine-to-Machine (M2M) — Lifecycle management, automated rotation, and repo leak risks

Scaling Access — Moving fast without compromising credential security

BAD | BARCELONA AI & DATA SYSTEMS

Scalable AI Infrastructure

Accelerate Operations

Centralize infrastructure to accelerate workflows and prevent administrative or technical bottlenecks.

Federated Control

Enforce robust security with federated SSO authentication coupled with automated policy enforcement.

Future-Proofing

Build a scalable foundation prepared for advanced multi-agent orchestrations and hybrid deployments.

Keyless-what?

Is this dark magic?

Static is a risk

Manual key rotation — Complicates operations

Key propagation — No control over where and who uses the keys

Key leak — Static long living secret can easily be stolen

BAD | BARCELONA AI & DATA SYSTEMS

User to Machine Architecture

01

One IDP & Source of Trust

Single identity provider serving as the ultimate source of truth for all users.

02

Universal JWT Tokens

Standardized JSON Web Tokens utilized across all platform services.

03

Tailored Virtual Keys

Dynamic virtual keys automatically provisioned to match exact user needs.

04

Built-in Observability

Turnkey automation and full telemetry benefits out of the box.

CLIENT ACCESS POINTS

Web UI Access

Interactive prompt testing, spend dashboards & usage logs

User Machine

CLI lite login | OpenCode

CORPORATE IDENTITY (IDP)

SSO – Web

OIDC standard redirect & MFA session cookies

SSO – Native

PKCE loopback authentication with short-lived tokens

AWS Account AI (Private VPC)

Zero Public Ingress

Private ALB

VPC Internal Load Balancer & TLS termination

EKS Cluster Workload

Pod Replica

LiteLLM Proxy

JWT ↔ Virtual Key Mapping

Validates OIDC claims (sub, email) & binds request to virtual budget without static keys

Postgres (Metadata)

Upstream LLM APIs

BAD | BARCELONA AI & DATA SYSTEMS

Machine to Machine Architecture

01

One IDP - GitHub

Single identity provider using GitHub as the primary source of truth for machine authentication.

02

JIT Access to Gateway

Just-In-Time dynamic access provisioned on demand for the LLM Gateway without static secrets.

03

Dedicated GitHub Action

Tailored workflow integration minting ephemeral OIDC tokens seamlessly within CI/CD pipelines.

04

Built-in Observability

Turnkey automation, monitoring, telemetry, and tracking benefits out of the box.

OIDC JWT Exchange | Ephemeral TTL Key | Auto-Decommission

GHA CI/CD PIPELINE

GHA Workflow

TRIGGER PHASE
Workflow requests GITHUB_TOKEN with id-token: write.

ACTION STEP
Custom action mints OIDC JWT using GitHub as IdP.

Job Completion Hook
Virtual key revoked; zero persistent secrets leaked.

GITHUB IDENTITY PROVIDER

GitHub OIDC IdP

token.actions.github...

Mint JWT: Claims include repo, branch, job ID

Public JWKS: Validated by LiteLLM

AWS Account AI (Private VPC)

CI/CD Private Ingress

Private ALB

Receives JWT & routes exchange

EKS LiteLLM Service

Auto-Expire

Validates GitHub IdP JWKS

Generates Ephemeral Virtual Key
Tied to repo claim with explicit duration (duration: 30m, budget: $20.00). Returned to runner.

Hard TTL Limit | Post-Job Revocation

BAD | BARCELONA AI & DATA SYSTEMS

Enabling AI at scale

Accelerate Adoption

Your everyday credentials grant you access to our LLM Gateway without having to manage separate credentials

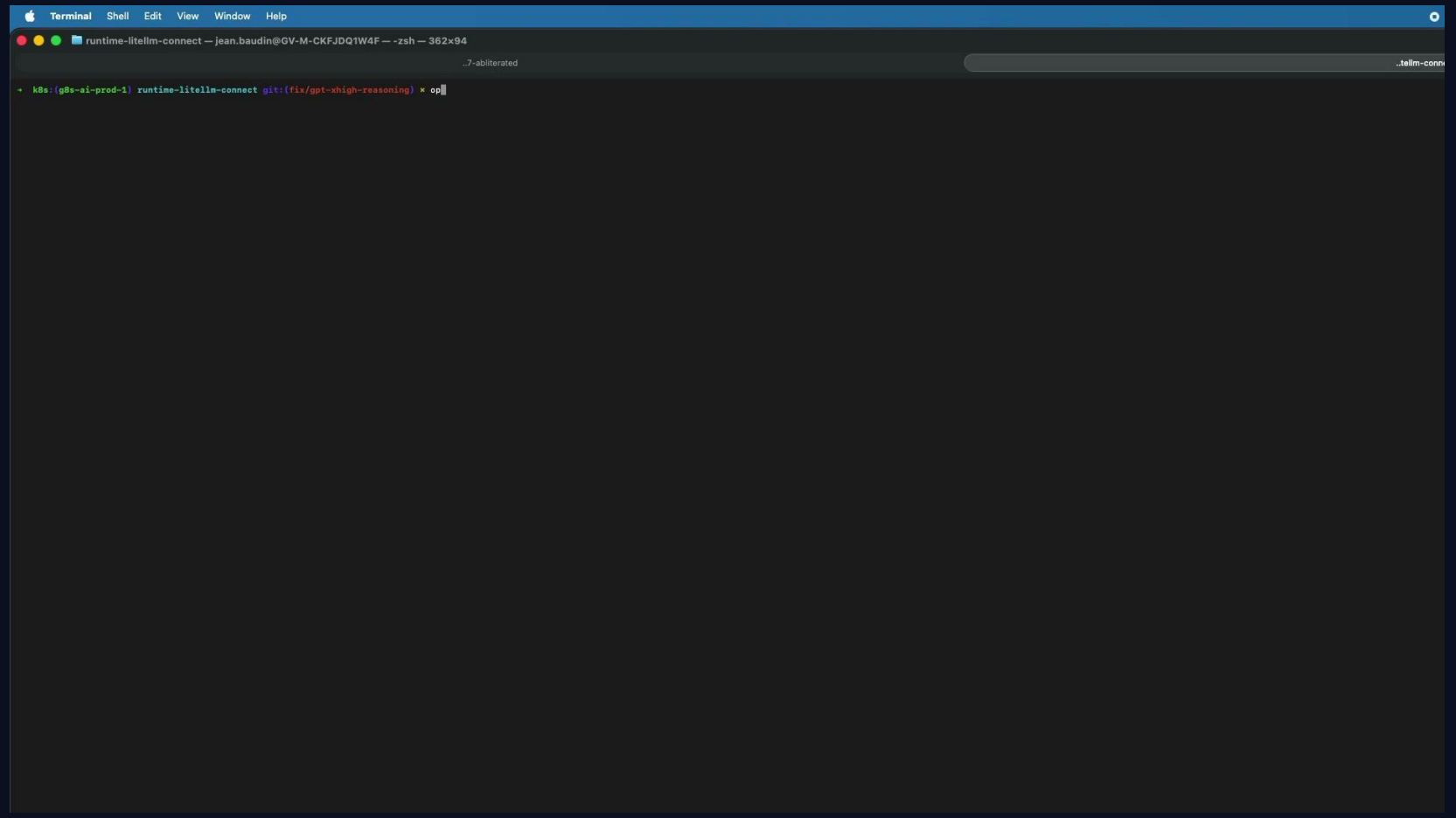
Increase security

JWT tokens are natively rotated. JWT tokens are tied to specific users. Offboarding makes token inactive. Better auditability.

Less operational overhead

No static secrets to provision. No need for custom automation to rotate keys or invalidate them. Less complications for security audits.

OpenCode Plugin & SSO



OpenCode Plugin

Short demo of the OpenCode plugin

Github Actions & LiteLLM

workflow.yaml · Your agentic workflow

```yaml
name: LiteLLM demo

permissions:
  contents: read
  id-token: write
  issues: write
  pull-requests: write

jobs:
  review:
    runs-on: glovo-arm64-shared-spot-medium
    steps:
      - name: Checkout pull request
        uses: actions/checkout@v6
        with:
          persist-credentials: true
          fetch-depth: 0

      - name: Exclude review output from Git
        run: printf 'review_comments.json\n' >> .git/info/exclude

      - name: Authenticate with LiteLLM
        id: litellm
        uses: Glovo/runtime-litellm-connect/apps/github-actions@main
        with:
          litellm-url: https://litellm.g8s-ai-stage.glovoint.com
          audience: https://litellm.glovoint.com/github-actions

      - name: Invoke qwen3-235b chat completion
        env:
          LITELLM_OIDC_TOKEN: ${{ steps.litellm.outputs.oidc-token }}
        run: |
          curl --fail-with-body --silent --show-error \
               --request POST \
               --url https://litellm.g8s-ai-stage.glovoint.com/v1/chat/completions \
               --header "Authorization: Bearer ${LITELLM_OIDC_TOKEN}" \
               --header "Content-Type: application/json" \
               --data '{
                  "model": "qwen3-235b",
                  "messages": [
                    {
                      "role": "user",
                      "content": "Reply with exactly: LiteLLM GitHub Actions test succeeded."
                    }
                  ],
                  "temperature": 0
                }'
```

BAD — Barcelona AI & Data Systems

BAD | BARCELONA AI & DATA SYSTEMS

Making security happy, without sacrificing DX


JAYRENSHAW
ANY
QUESTIONS?

TALK

BAD | BARCELONA AI & DATA SYSTEMS

Human Gaze as an Alignment Signal for LLMs and VLLMs

Eye-tracking signals make LLM and VLLM alignment more human — better reward models, fewer hallucinations.

Ángela López-Cardona (Researcher), **Carlos Segura** (Sr. Researcher) · Telefónica Innovación Digital

BAD | BARCELONA AI & DATA SYSTEMS

What we will cover

Introduction & Background

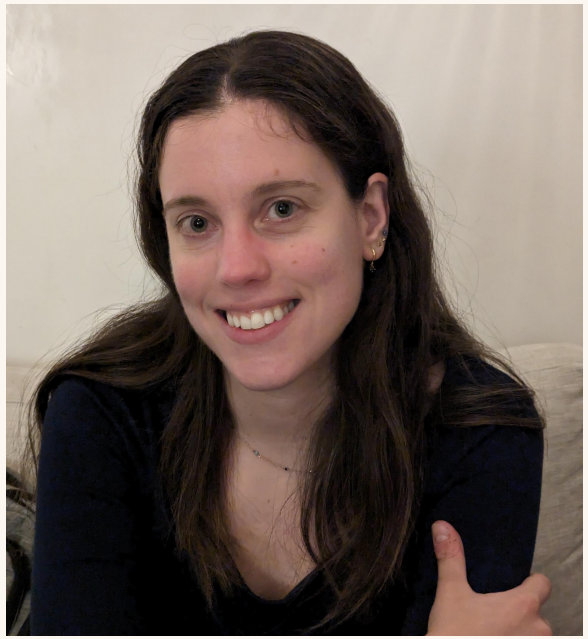
GazeReward

GazeSteer

Conclusions

BAD | BARCELONA AI & DATA SYSTEMS

Human Gaze as an Alignment Signal



Ángela López-Cardona

Researcher, Telefónica Innovación Digital · Industrial PhD candidate, UPC

- Works on the integration of cognitive signals into reward modeling, alignment, and interpretability.
- Holds Master's degrees in Industrial Engineering and Artificial Intelligence.

BAD | BARCELONA AI & DATA SYSTEMS

Human Gaze as an Alignment Signal

Jesus Omaña Iglesias

Lead Researcher, Telefónica Innovación Digital

- Research leader in distributed systems and AI across telecom, edge, and cloud. Expertise spanning federated learning, mobility-aware AI, RAN optimization, orchestration, and sustainable infrastructure for large-scale intelligent systems.

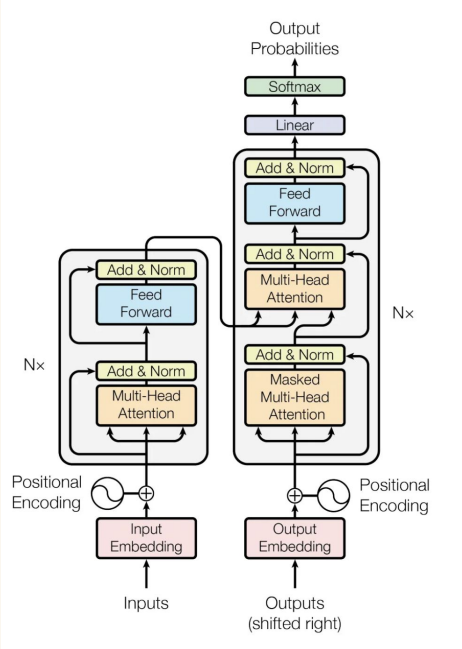
PART 1

Introduction & Background

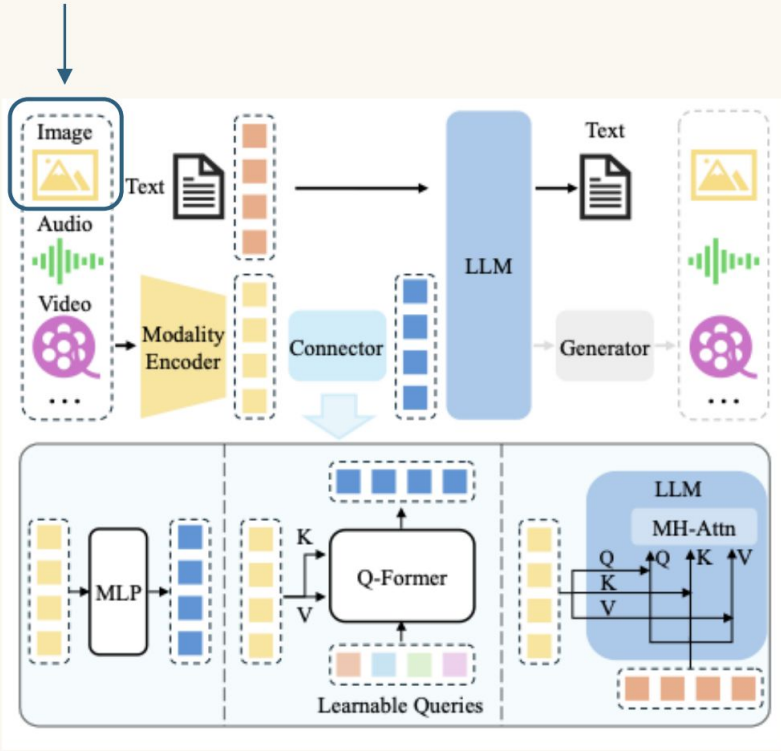
Why human attention matters for alignment

Large Language Models (LLMs)

The origin of these models was in 2017, when Google Brain published the paper called Attention Is All You Need [43].



Transformer architecture



Typical Multimodal LLM (MLLM) architecture [85]

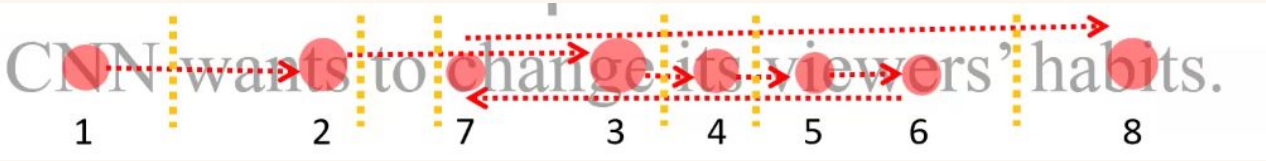
LLMs challenges

- **Scarcity** of sufficient data or the suboptimal use of existing data collections for training, which creates a significant bottleneck [12].
- **Environmental cost** associated with training large-scale models resulting in substantial CO_2 emissions [42, 29, 23]
- **Aligning models** with human values, such as the ability to follow instructions, avoid harmful behaviors, and engage in safe, user-friendly interactions [53, 7].
- **Models hallucinating,** which leads to the generation of factually incorrect content [50, 7].

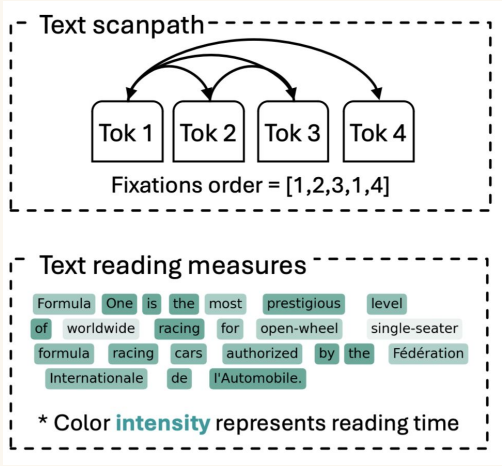
Eye tracking

- Eye-tracking devices gather **raw data** about scanpaths, frequency and duration of fixations, as well as pupillary dilation variations.
- Algorithms to detect **fixations sequence**.
- From this **fixations sequence** → reading measures.

Acronym	Measure
FFD	First Fixation Duration
GPT	Go-Past Time
TRT	Total Reading Time
nFix	Number of Fixations
fixProp	Proportion of participants



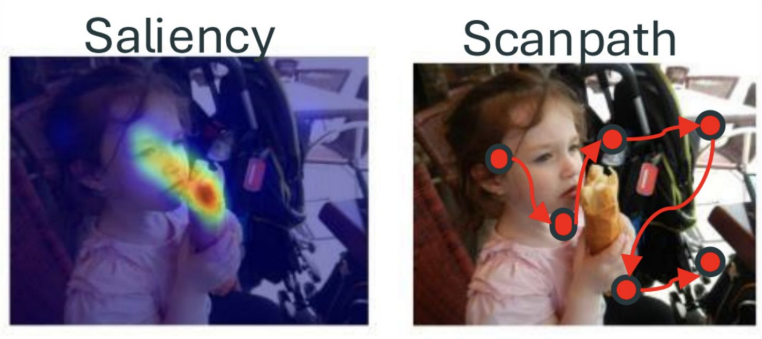
Fixation sequence over a sentence, including regressions.



Eye tracking

- Eye-tracking devices gather **raw data** about scanpaths, frequency and duration of fixations, as well as pupillary dilation variations.
- Algorithms to detect **fixations sequence**.
- From this **fixations sequence** → reading measures.

Acronym	Measure
FFD	First Fixation Duration
GPT	Go-Past Time
TRT	Total Reading Time
nFix	Number of Fixations
fixProp	Proportion of participants



Saliency map and scanpath recorded on an image.

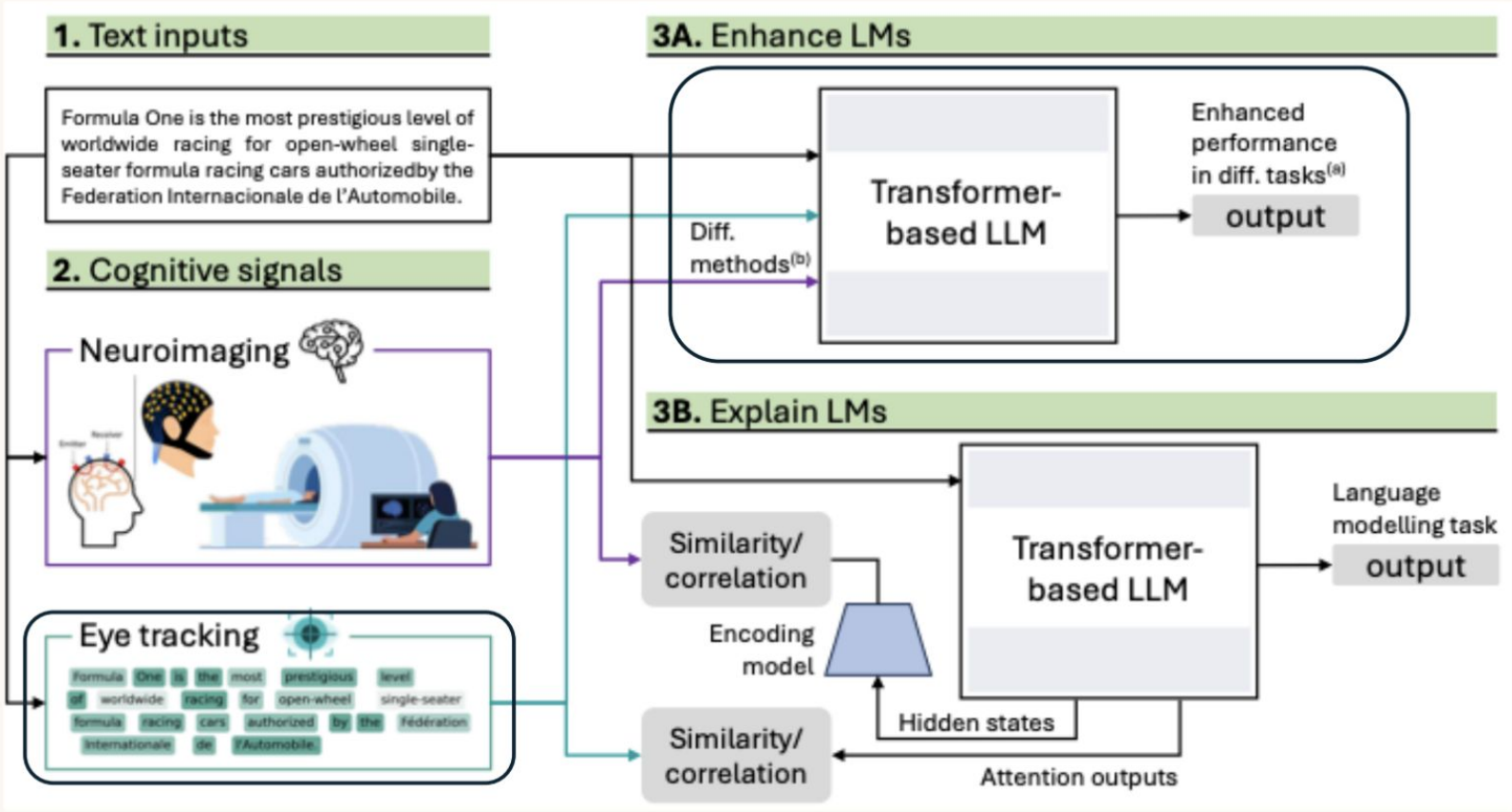


Figure 1. Schematic overview illustrating the use of cognitive signals (2) to enhance LLMs (3A) and to explore similar representations (3B).

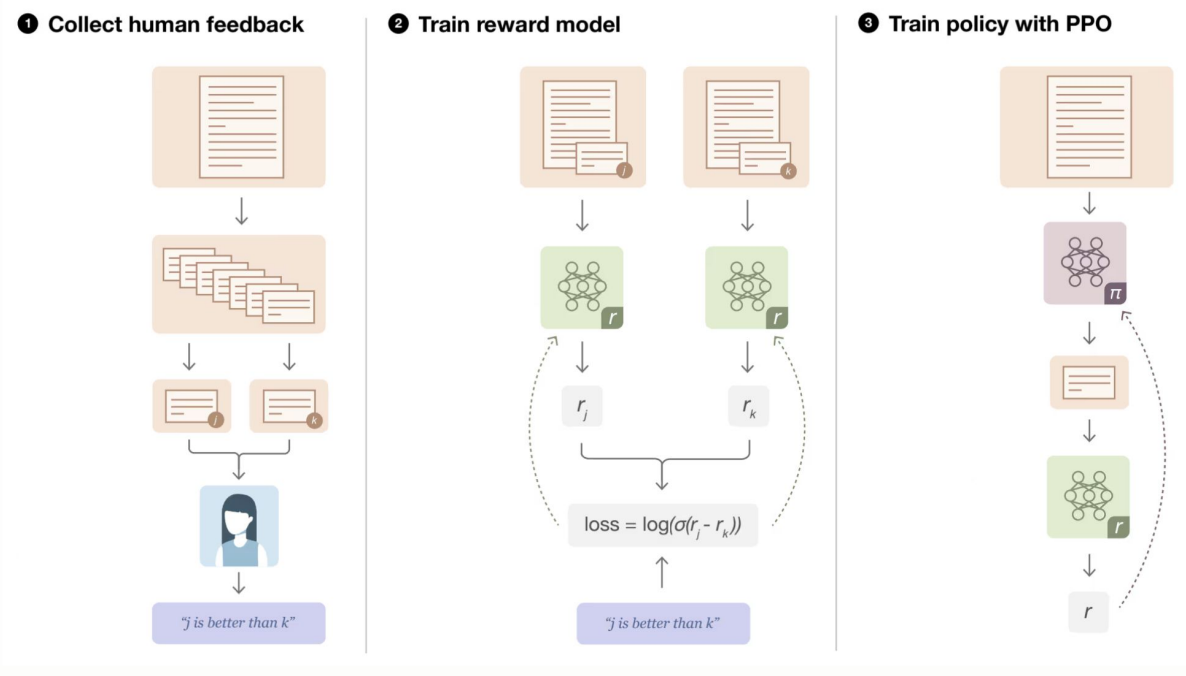
PART 2

GazeReward

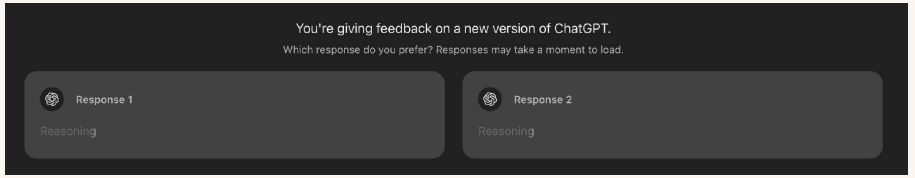
Gaze-based rewards for large language models

Human alignment with RLHF

- Unsupervised pretraining from raw text → **Language understanding**
- Large scale instruction tuning and reinforcement learning → **Human alignment**



- **RLHF phases:** (1) collecting feedback, (2) training a Reward Model (RM), (3) optimising the LLM using RL (PPO).

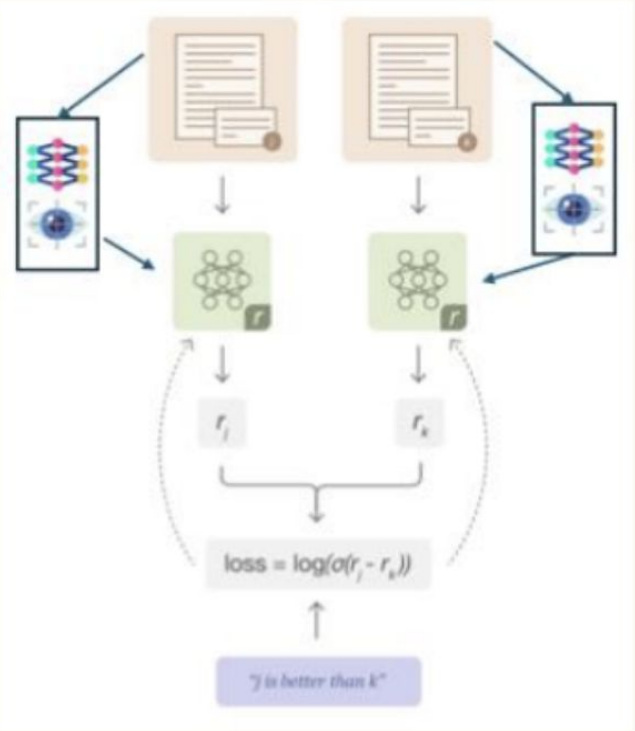


Human preference feedback interface.

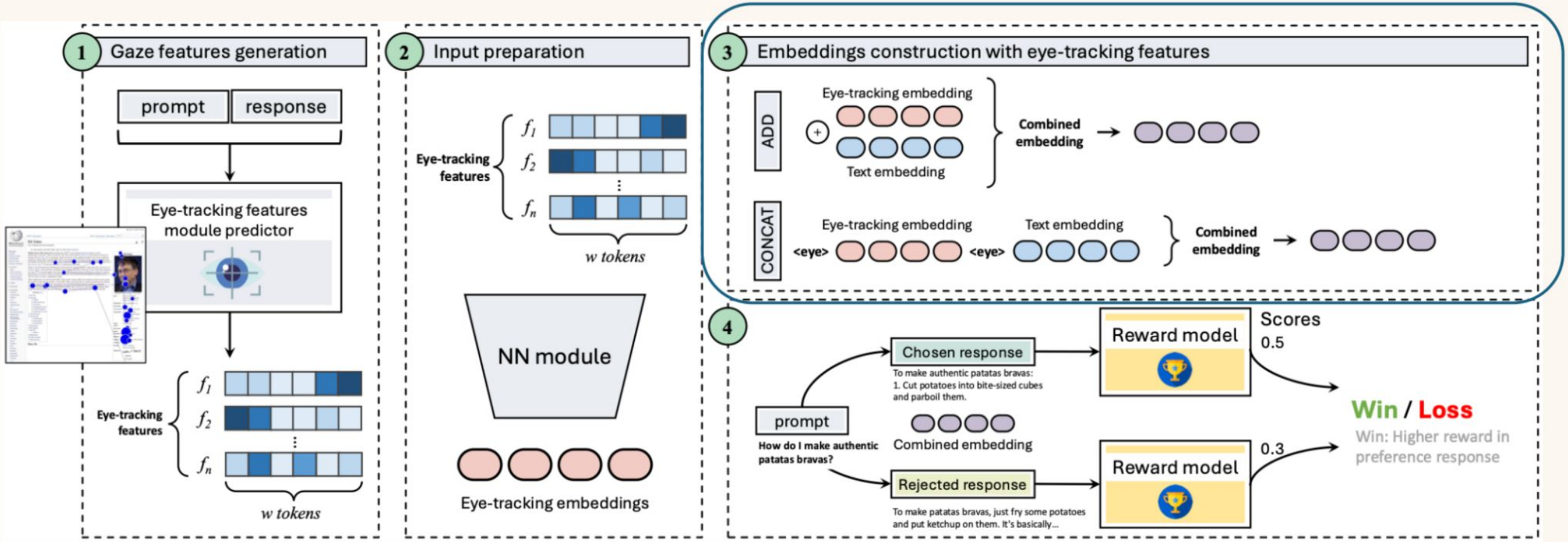
Seeing Eye to AI: Human Alignment via Gaze-Based Response Rewards for Large Language Models

ICLR 2025

- **Experiments**
- 2 inclusion methods: GazeConcat, GazeAdd.
- 2 ET generative models.
- 3 ET features combinations.
- 3 open source backbone models.
- 2 datasets.



Introduction & Method



Results

Reward modeling accuracy (%) — OASST1

		Llama-3-8B-Instruct		Llama-3-8B		Mistral-7B	
	baseline	65.9 ± 0.5	diff (%)	65.5 ± 2.1	diff (%)	66.3 ± 0.1	diff (%)
GazeConcat	$fcomb_1$	$69.0 \pm 0.4*$	4.7	69.3 ± 0.6	5.9	67.6 ± 1.7	2.1
	$fcomb_{2.5}$	$70.2 \pm 0.3**$	6.6	$\mathbf{71.5 \pm 0.5}$	9.2	$70.2 \pm 0.4*$	5.9
	$fcomb_{2.2}$	$70.0 \pm 0.4**$	6.3	71.2 ± 0.8	8.8	$\mathbf{71.0 \pm 1.0}$	7.1
GazeAdd	$fcomb_1$	68.9 ± 0.9	4.6	68.9 ± 1.0	5.3	-	
	$fcomb_{2.5}$	$\mathbf{70.2 \pm 0.1}*$	6.6	69.5 ± 0.3	6.1	-	
	$fcomb_{2.2}$	$69.0 \pm 0.4*$	4.7	68.3 ± 0.7	4.4	-	

Reward modeling accuracy (%) — HelpSteer2

		Llama-3-8B-Instruct		Llama-3-8B		Mistral-7B	
	baseline	54.7 ± 0.7	diff (%)	53.3 ± 0.8	diff (%)	54.1 ± 0.3	diff (%)
GazeConcat	$fcomb_1$	$61.1 \pm 1.2*$	11.8	$59.1 \pm 0.2*$	10.8	57.6 ± 2.3	6.5
	$fcomb_{2.5}$	58.5 ± 1.6	7.0	$60.3 \pm 0.5**$	13.2	$\mathbf{58.7 \pm 2.4}$	8.5
	$fcomb_{2.2}$	60.6 ± 3.3	10.9	57.9 ± 2.0	8.6	56.0 ± 2.4	3.4
GazeAdd	$fcomb_1$	$\mathbf{62.3 \pm 0.6}**$	13.9	$\mathbf{62.4 \pm 1.0}**$	17.0	-	
	$fcomb_{2.5}$	$59.6 \pm 1.1*$	9.0	$58.6 \pm 1.2*$	10.0	-	
	$fcomb_{2.2}$	$60.3 \pm 0.5**$	10.2	$59.3 \pm 0.1*$	11.3	-	

Conclusions

- 👁️ **Implicit Feedback for Reward Models:** a new framework integrates implicit feedback into the RM, improving alignment and synthetic data generation.
- 📚 Validated on **open-source** LLaMA-3 and Mistral models, with consistent gains in preference prediction and scalability through **model-generated** eye-tracking features.
- 🔗 Limited to small, English-only datasets; future work targets **larger models, broader data collection** and integration into **RLHF** or rejection-sampling pipelines for improved LLM alignment

Karim Galliamov, Titov & Pershin — EMNLP 2025

(a) Validation scores gained by policy in PPO

(b) Training rewards gained by policy in PPO

PART 3

GazeSteer

Steering visual attention with human saliency

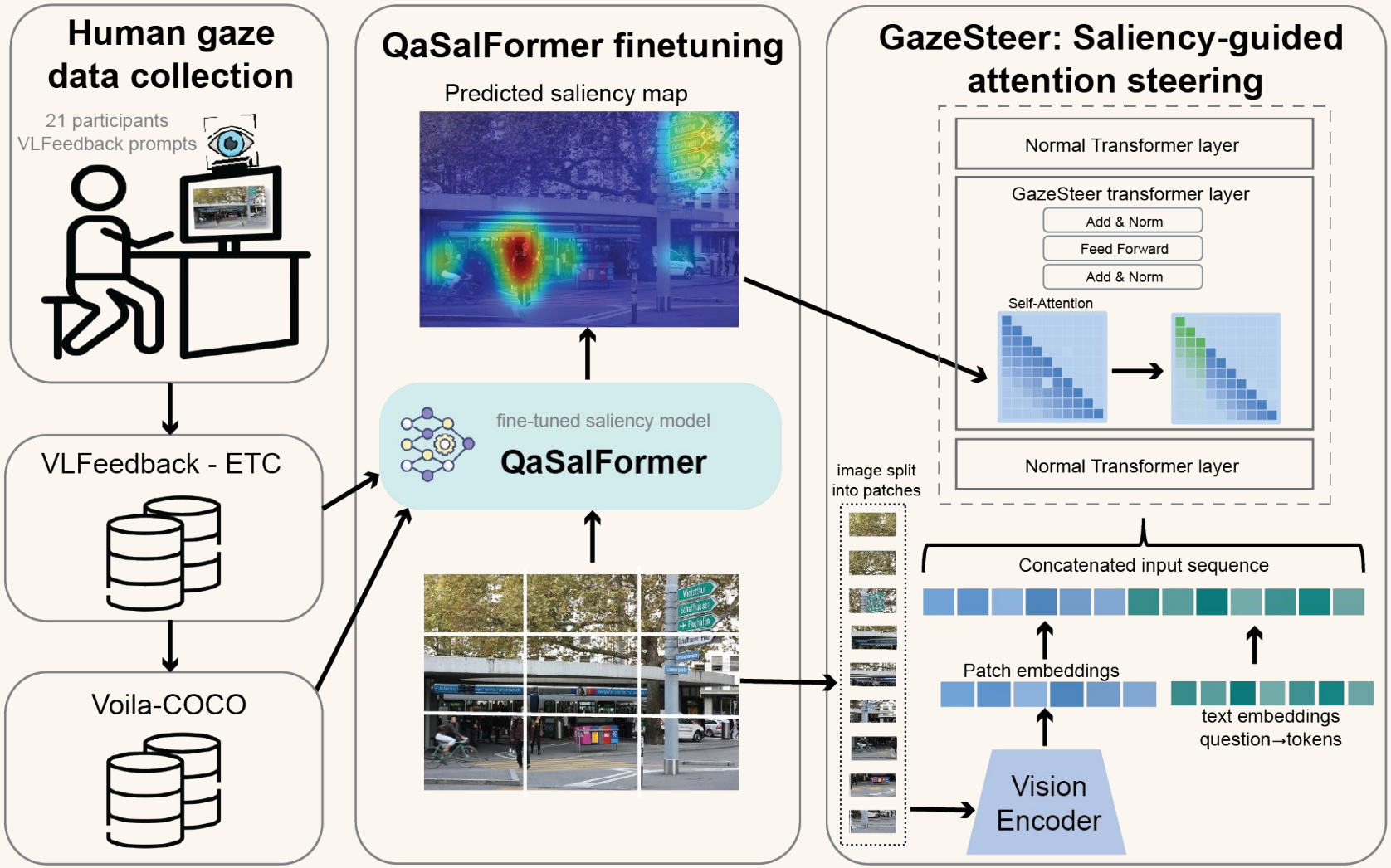
Look Where Humans Look: Steering VLLM Attention with Question-Conditioned Human Saliency to Mitigate Hallucinations

Under Review

- **VLLMs hallucinate objects** — captions mention what is not in the image.
- **Attention drifts** — models often attend to regions humans never look at.
- **Human saliency is a prior** — question-conditioned gaze shows where the answer lives.
- **GazeSteer** — inject predicted saliency into attention at inference, no retraining needed.

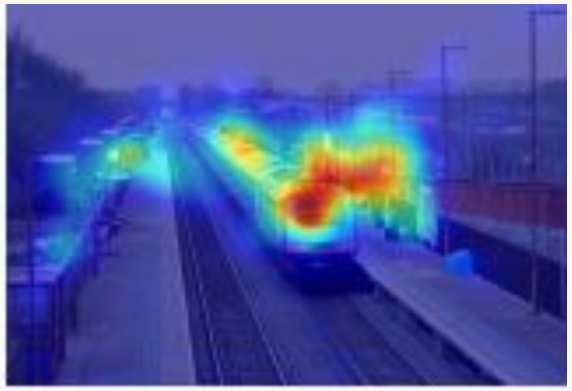
BAD | BARCELONA AI & DATA SYSTEMS

Method



Hallucination mitigation

Reducing hallucinations in Visual LLMs by leveraging visual saliency — we developed QSalFormer, a model for predicting saliency in this context.



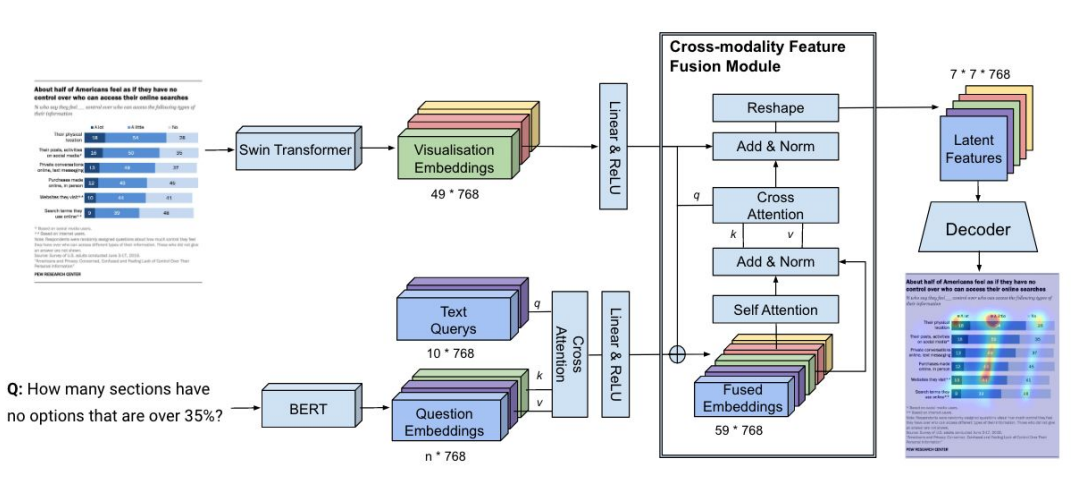
(a) Human saliency

(b) LLaVA-7B attention

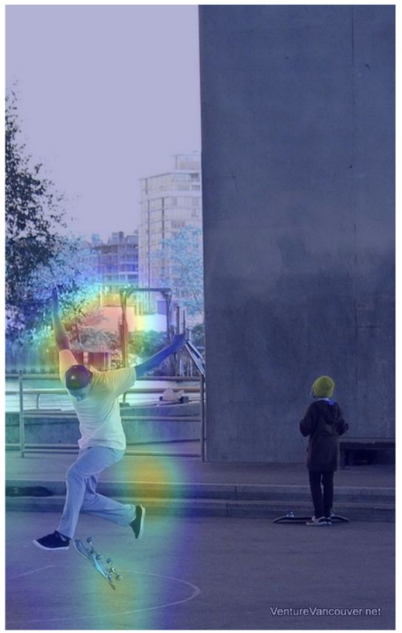
(c) LLaVA-13B attention

Human saliency and LLaVA-7B/13B attention maps for prompt 8: What are the prominent objects in the image?

Saliency model & Dataset



VisSalFormer — cross-modality feature fusion architecture

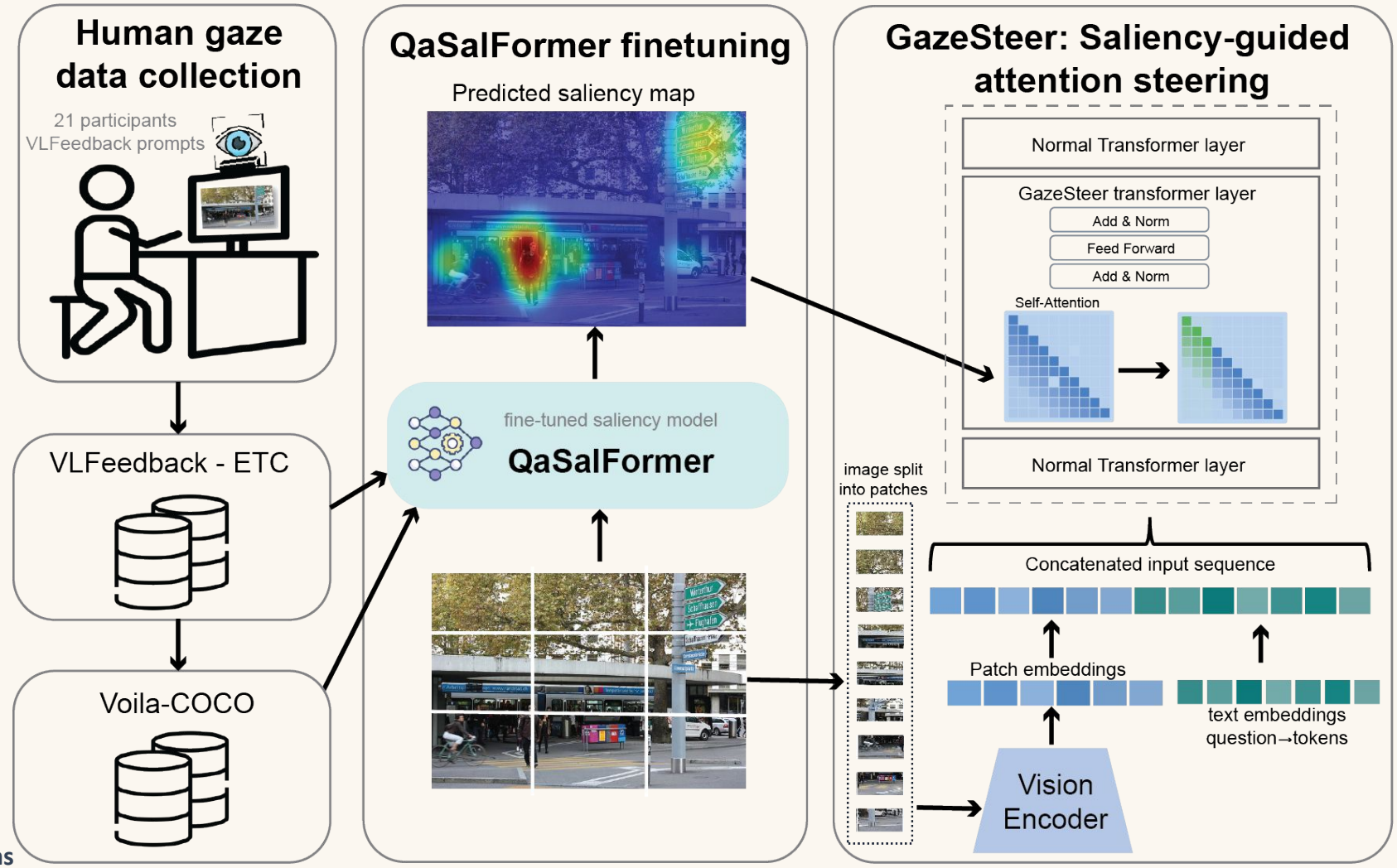


Caption: Bottom left side of the image a man is doing skateboarding and jumping.
Question: What is the skateboarder in the bottom left doing?
Indirect Question: What is he doing?
Answer: The skateboarder in the bottom left is performing a trick, jumping into the air with his skateboard.

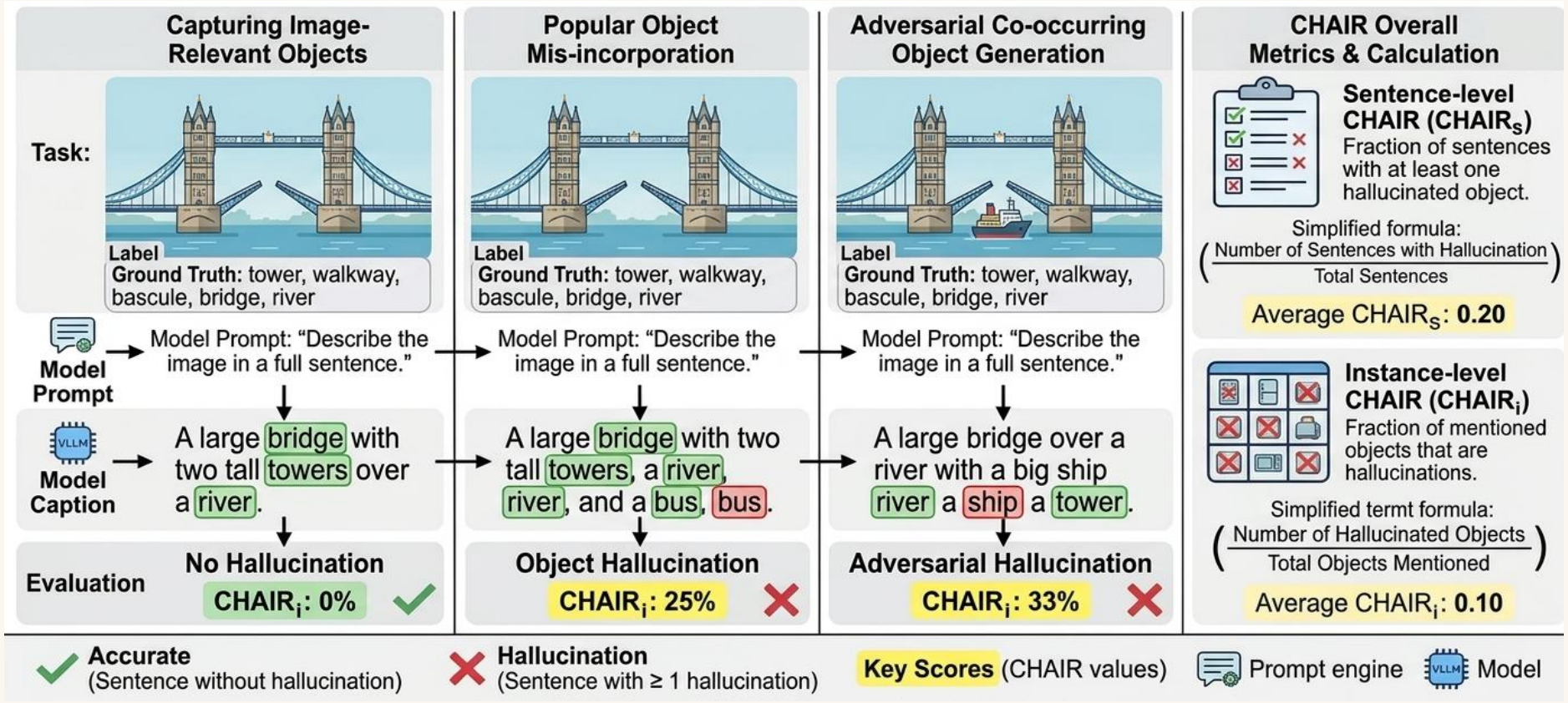
Voila-COCO dataset [48]

BAD | BARCELONA AI & DATA SYSTEMS

Method



CHAIR: Measuring Caption Hallucination through Image Relevance



Results: saliency steering reduces hallucination

			LLaVA-1.5-7B				LLaVA-1.5-13B			
Method	α	β	$C_S \downarrow$	$C_I \downarrow$	F1↑	Len	$C_S \downarrow$	$C_I \downarrow$	F1↑	Len
Greedy†	–	–	53.6	14.0	76.5	101	48.6	12.2	**77.9**	102
Mean	0.5	0	26.0	6.3	<u>76.7</u>	91	32.8	9.4	<u>76.7</u>	91
Saliency	0	0.65	26.4	7.2	73.8	74	28.0	9.5	73.6	77
Saliency (random)‡	0	0.65	35.6	9.0	**77.8**	89	37.6	10.9	76.5	98
Saliency	0	0.68	<u>20.4</u>	**5.8**	71.0	65	**22.8**	9.8	71.4	67
Mean + saliency	0.3	0.35	20.4	<u>5.9</u>	71.0	63	25.6	<u>8.6</u>	73.7	75
Mean + saliency	0.5	0.04	**19.6**	6.3	73.0	77	<u>24.8</u>	**7.8**	74.3	86

CHAIR hallucination evaluation and F1 on LLaVA-1.5-7B/13B (layers 5–19, prompt “What is in the image?”). α: mean strength; β: saliency strength. † greedy decoding; ‡ random Gaussian saliency (control).

TAKEAWAY

Where people look tells us what matters — and models align better when they look there too.

Contact
angela.lopezcardona@telefonica.com

Ángela López Cardona
AI Researcher at Telefónica | PhD Candidate at UPC

Acknowledgements — with the support of:

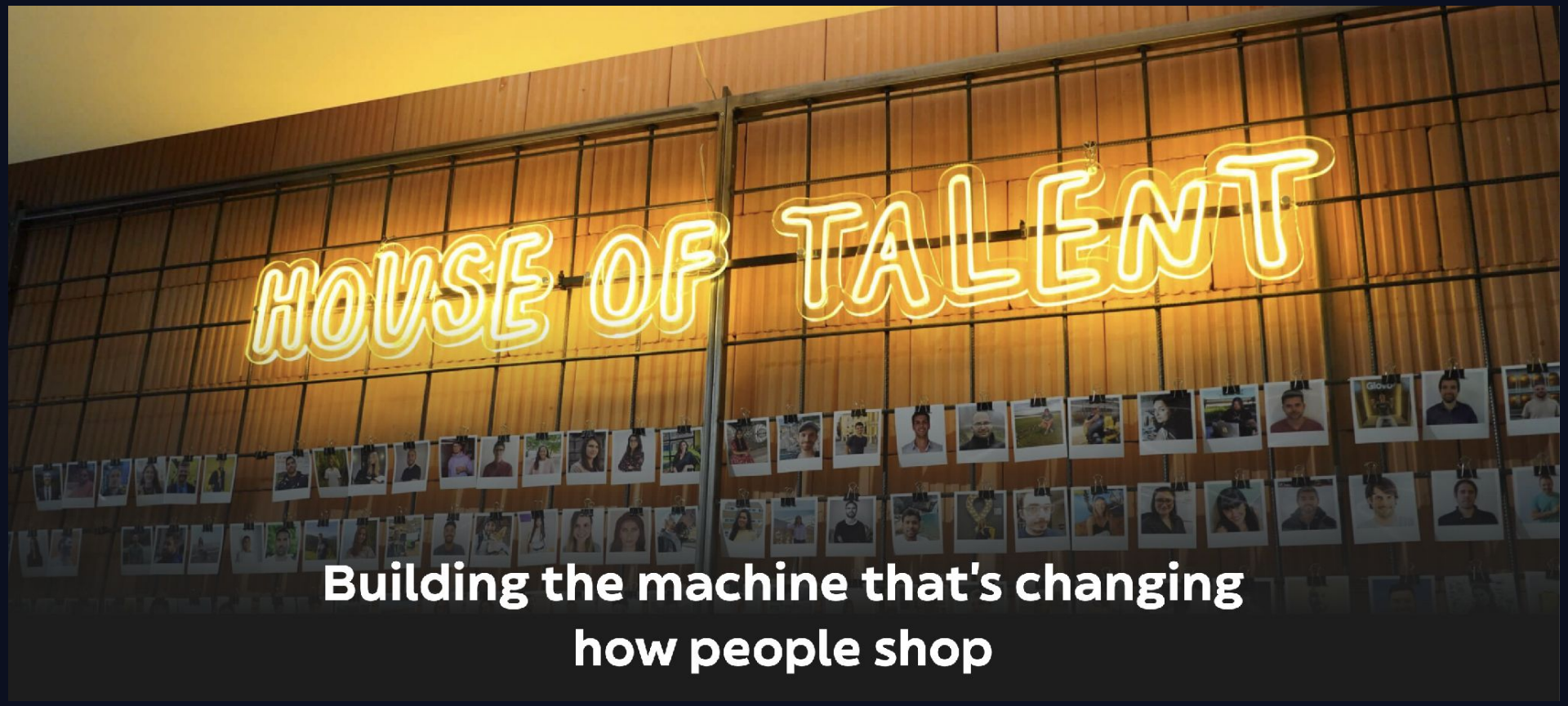
BAD | BARCELONA AI & DATA SYSTEMS

COMMERCIAL BREAK

We’re Hiring!

GLOVO

BAD | BARCELONA AI & DATA SYSTEMS



careers.glovoapp.com

GLOVO

BAD | BARCELONA AI & DATA SYSTEMS

Ready for the “Ride of Your Life”? Join our Talent House!

- **Data:** Jr. Analytics Engineer - AI & Data Enablement
- **Security:** Cyber Defence & Incident Response Engineer
- Multiple **Early Careers Booster** and **Internship** opportunities

careers.glovoapp.com/tech-glovo/

TALK

BAD | BARCELONA AI & DATA SYSTEMS

Towards an AI software Factory for Data Systems

From coding assistance to end-to-end, self-improving software development

Laura Pereira Sanchez (Senior Applied Scientist), **Subru Krishnan,** (Principal Architect) · Microsoft

Laura Pereira Sánchez

Senior Applied Scientist

Brings large-scale data analysis from particle physics into post-training LLMs that make Copilot better across Azure data products.

Subru Krishnan

Principal Architect

Translates systems research into production technologies for cloud-scale data infrastructure and autonomous data services.

BAD | BARCELONA AI & DATA SYSTEMS

AI speeds up coding, but not the end-to-end SDLC

Amdahl's Law: once coding accelerates, the remaining lifecycle becomes the bottleneck.

TARGET	CODE	REVIEW	OPS
Choose valuable work	Generate and optimize	Validate safely	Operate and respond

Focus **Data Systems** + **Evolutionary Coding Tasks** measurable objectives enable automatic feedback

Why coding became an ideal domain for AI

Software engineering already contains both rich decision traces and automatic feedback.

DATA

- Documentation
- Code history
- PR discussions
- Design decisions

Provides high-quality training traces

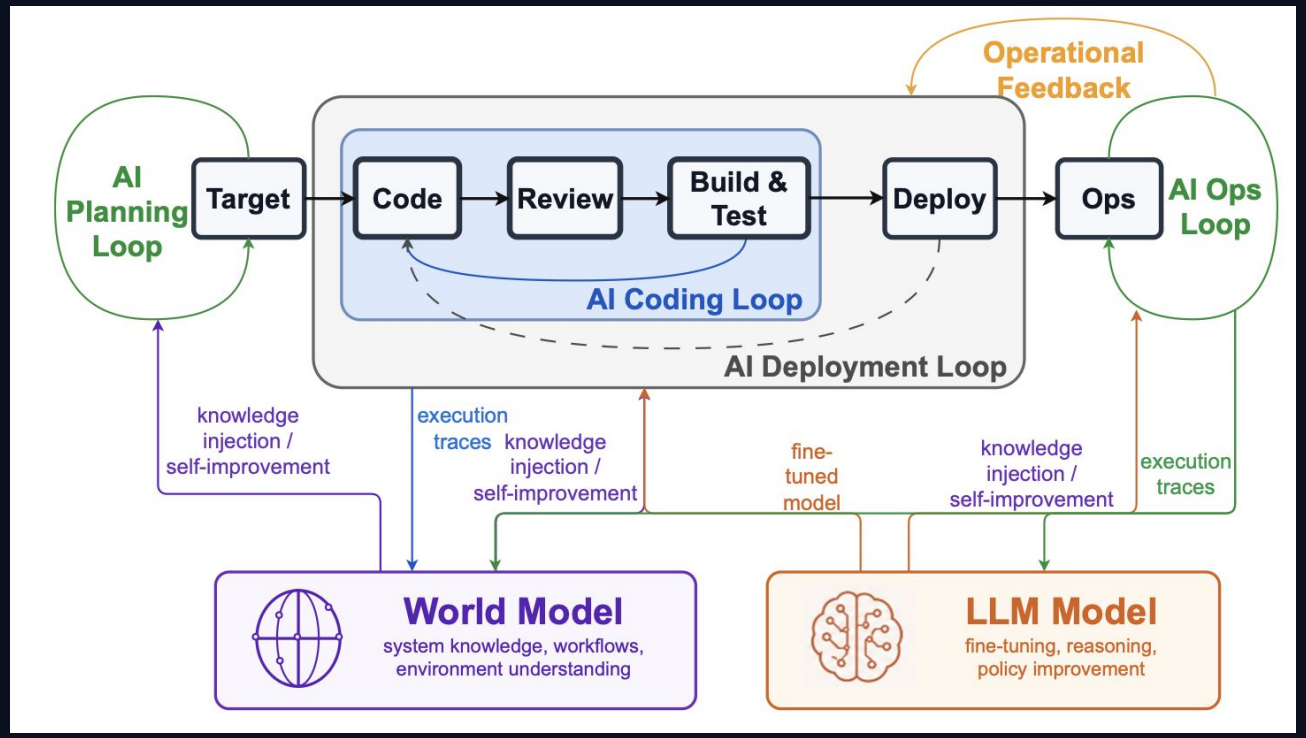
FEEDBACK

- Compilers
- Tests and benchmarks
- Linters
- Runtime signals

Enables agentic loops and RL

The AI Software Factory extends this data + feedback pattern across the whole SDLC.

A vision for an AI-enhanced software life cycle



AI coding is just one step towards accelerating the SDLC

- AI planning loop chooses best way to act
- Self-improving World Model gives detailed data and grounds the LLM
- Agentic coding loop writes, reviews, test and deploys code
- Ops and execution traces are used as feedback to the coding loop
- All these traces can be used to fine-tune the LLM model and improve the world model!

Building blocks

World Model

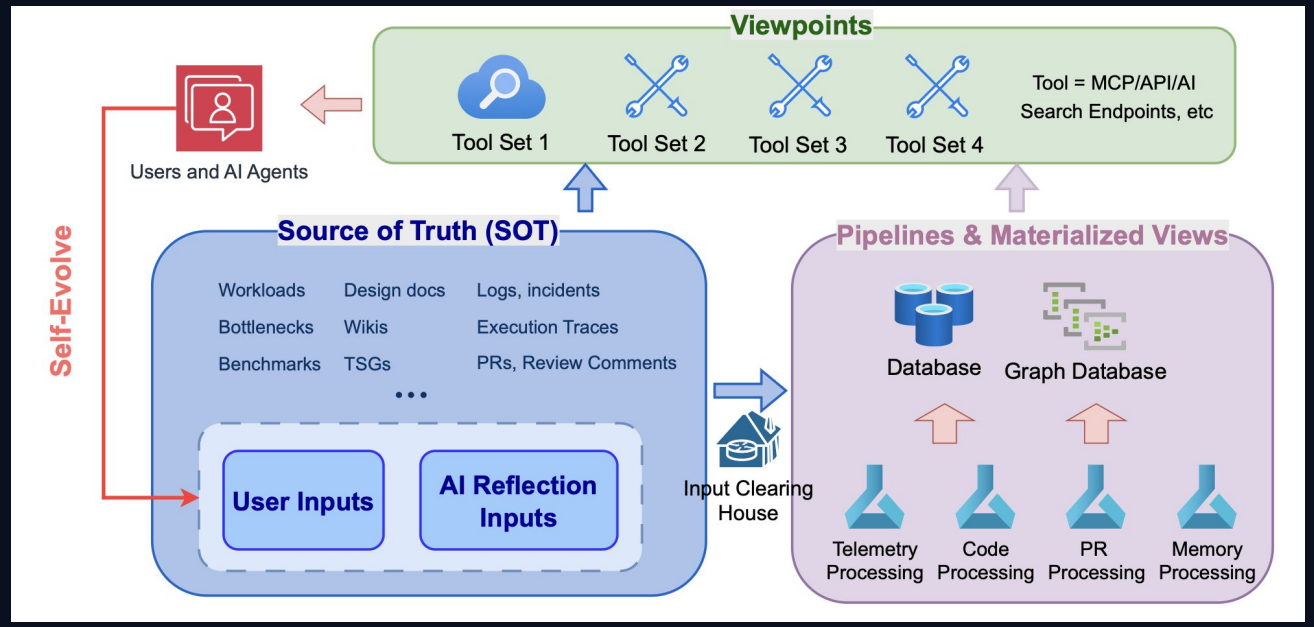
Targeting Agent

Darwin: Evolutionary Agentic Swarms

AutoSRE: Improving Site Reliability Engineering with Automated Ops

Evolution of the LLM: Fine-tuning Loop

World Model Architecture



Grounding the LLM with knowledge

- The world model handles the missing data problems AI face in the SDLC.
- It grounds the knowledge by providing access to SOT.
- Viewpoints allow to test and modify the predictions from AI.
- New user inputs + AI reflections are used to self-evolve the world model.

Targeting Agent: Hotspan agent harness

Determines what to change (target) and how to measure its success

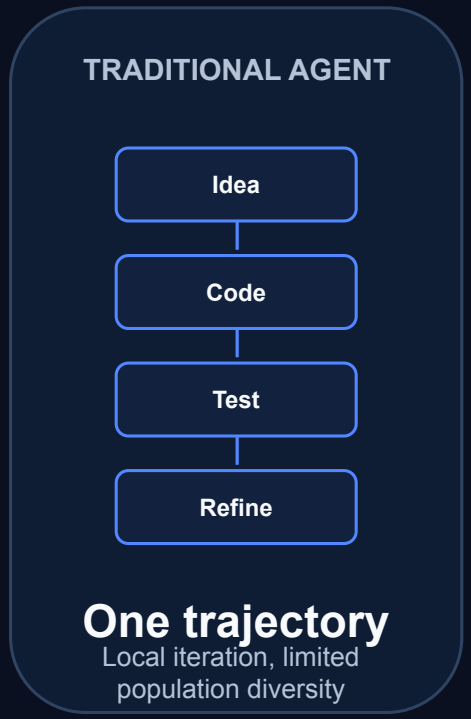
Leverages provenance and evidence grounded in the world model

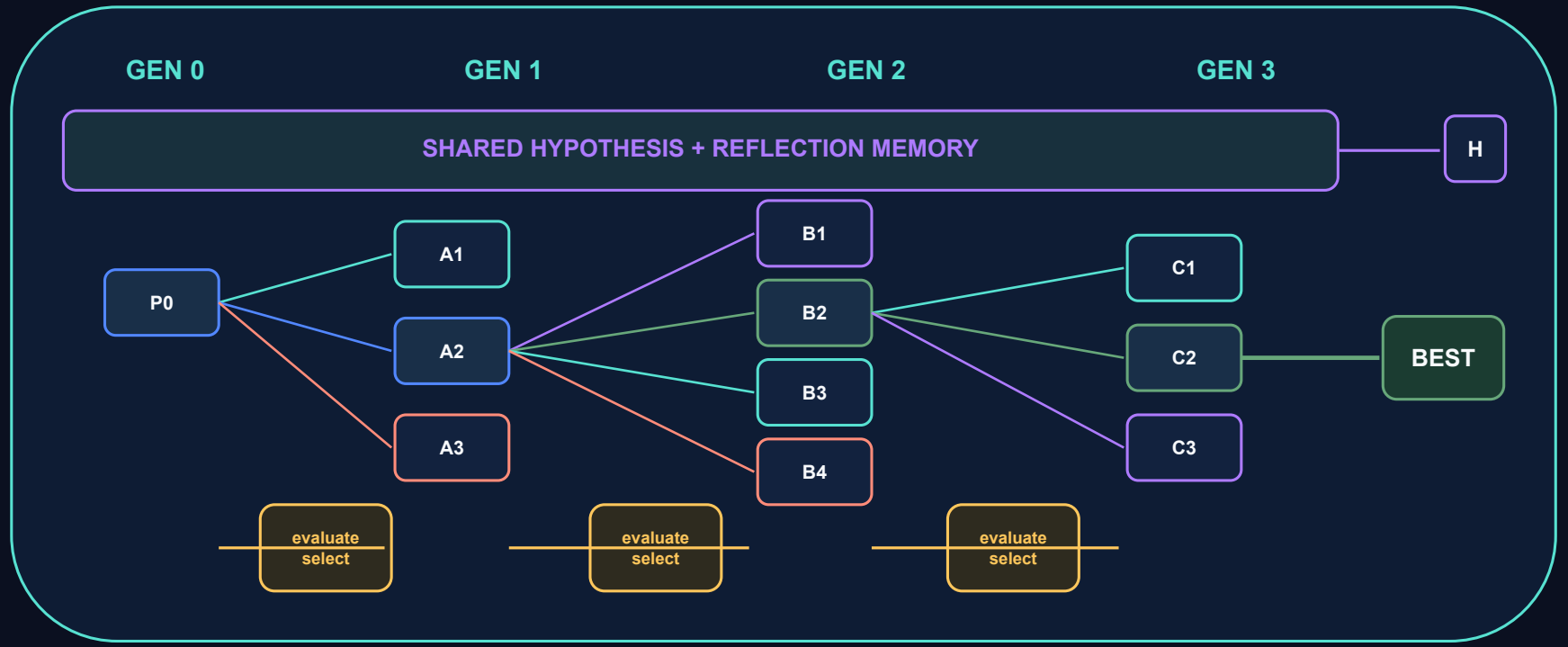
- Step 1) **Transforms ambiguous optimization goals** i.e. “make it faster” **into actionable objectives** by generating true target metrics (e.g. query latency) and identifying constraints (e.g. throughput cost)
- Step 2) **Gather evidence from world model** and **produces multiple targets**. Hotspan queries the world model viewpoints to extract telemetry, code analysis and code review.
- Step 3) **Grounded prioritization** via expected return on investment (ROI)
- Step 4) **Packages** each selected **target as** an **Evolutionary Coding Task** to be optimized by **Darwin**

BAD | BARCELONA AI & DATA SYSTEMS

A swarm evolves a population across generations

Parallel agents explore each generation; measured selection seeds the next generation and shared memory guides future hypotheses.





PARALLEL WITHIN A GENERATION

Many agents test different hypotheses at once

SELECTION BETWEEN GENERATIONS

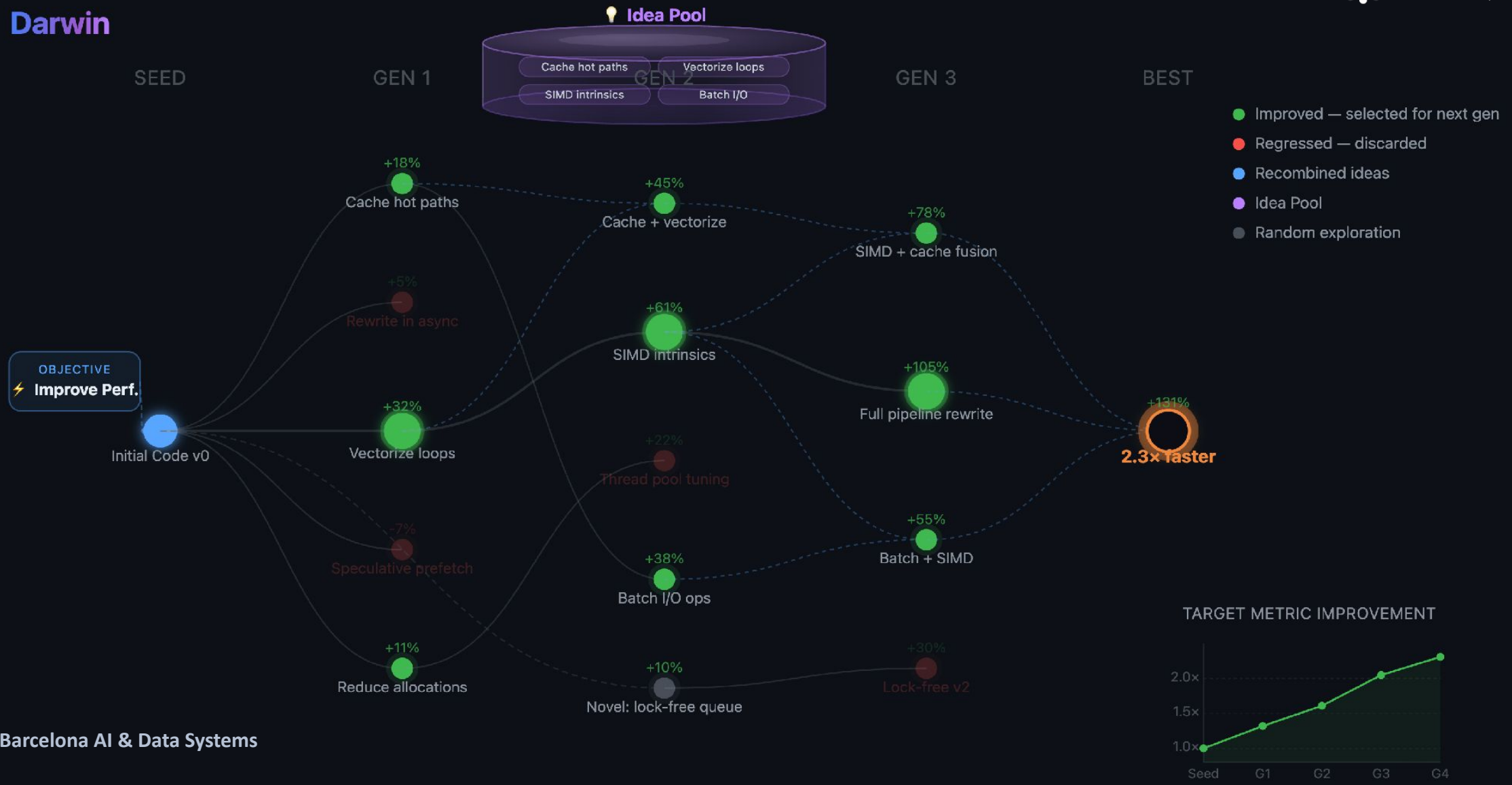
Only measured candidates seed the next population

MEMORY ACROSS GENERATIONS

Reflections guide new hypotheses and avoid repeats

BAD | BARCELONA AI & DATA SYSTEMS

Darwin



Darwin Impact on OSS and Microsoft

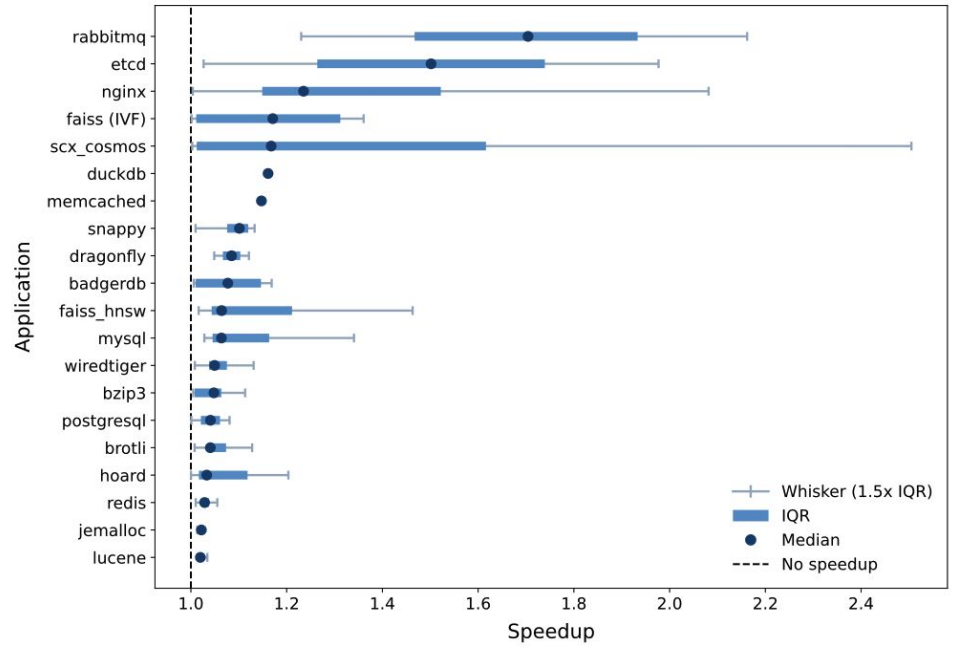


Figure 4: Darwin improvements of OSS projects.

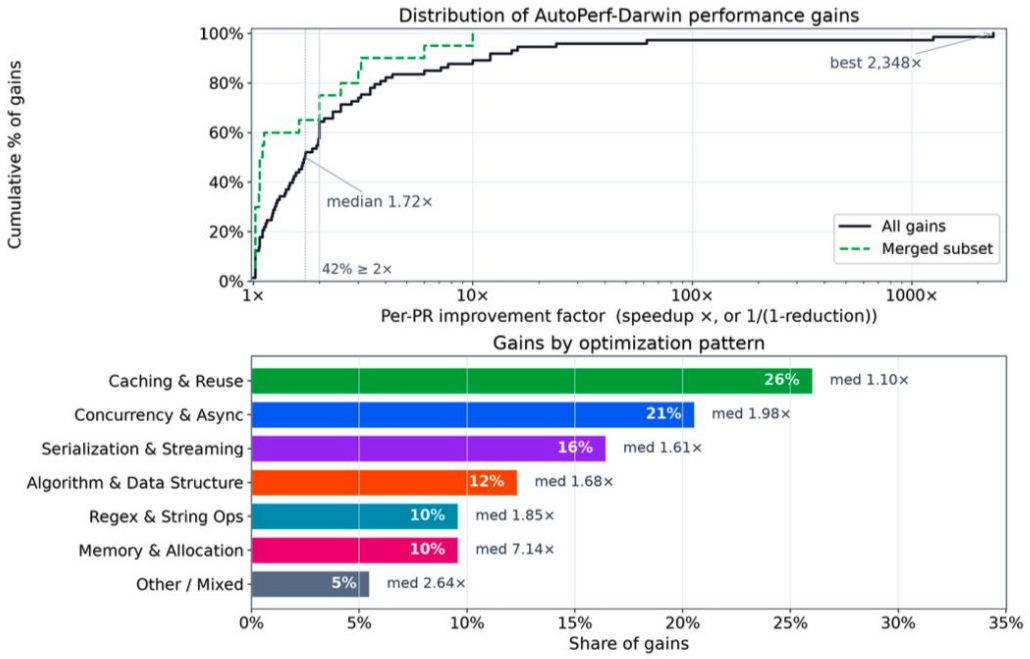


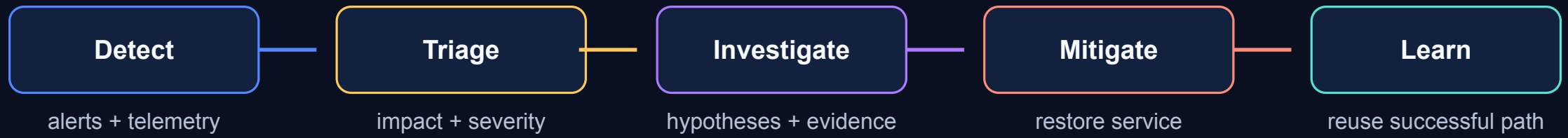
Figure 5: Production Evidence for Autoperf.

$100K public-token cost → **45/45** OSS projects improved → **Top 20: 13.9%** average gain
10% deployment footprint → **$200M–$600M/year** estimated COGS savings

BAD | BARCELONA AI & DATA SYSTEMS

Faster development requires faster incident response

Some failures only appear in production due to scale, diverse workloads, or cloud conditions.



AutoSRE An **Ops-specific World Model** and **agent loop** for incident detection, triage, root-cause analysis, and mitigation.

Goal: improve the AI-to-human toil ratio while preserving safe, evidence-grounded operations.

BAD | BARCELONA AI & DATA SYSTEMS

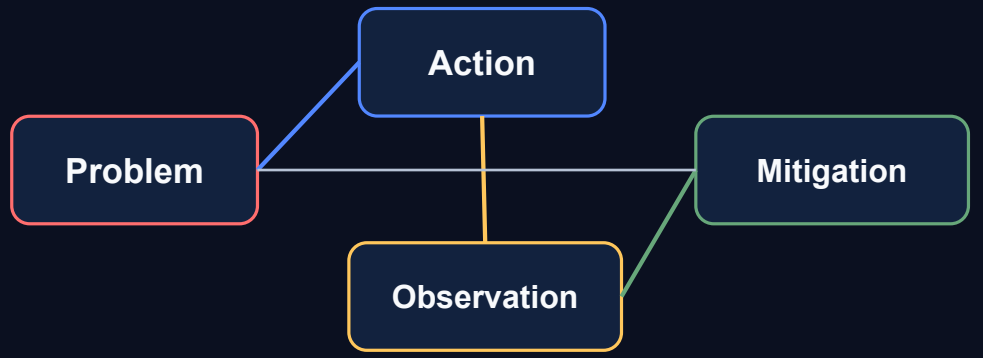
Resolved incidents become reusable investigation graphs

Agents retrieve high-evidence paths instead of reading the entire incident history.

Operational evidence

- Incident tickets
- Alerts + telemetry
- Troubleshooting guides
- Mitigation histories
- Investigation discussions

Incident-investigation graph



Problem → Action → Observation

Live incident

1. Match symptoms
2. Traverse high-evidence paths
3. Retrieve proven actions
4. Compare expected observations

BAD | BARCELONA AI & DATA SYSTEMS

Evaluation: Online Traversal

Graph-guided vs RAG baseline on ~93 incidents, with the graph learned from1,300 incidents

Evaluation Metrics

Reach Mitigation — reaches a mitigation action (binary)
Groundedness — conclusions supported by diagnostic queries (binary)
Usefulness — LLM-judged score (1–5) vs human resolution logs
Avg. KQL Succ. — # successfully executed KQL queries per incident
KQL Tbl. Recall — % of human-logged Kusto tables covered
Action Recall — % of human actions reproduced by system

Held-Out Results (93 incidents)

	SuperDRI	RAG
Reach Mitigation	98.5%	90.9%
Groundedness	96.2%	77.3%
Usefulness	3.49	3.32
Avg. KQL Succ.	10.0	2.5

Expert Review (19 blind evals)

Metric	SuperDRI	RAG
Avg. Score (1–5)	4.95	3.68
Perfect 5s	94.7%	26.3%

BAD | BARCELONA AI & DATA SYSTEMS

The factory learns at two different layers

Not every lesson belongs in model weights.

Evolve the World Model

For knowledge that is:

- Temporary
- Environment-specific
- Revisable
- Provenance-sensitive

Examples: workflows, incidents, dependencies

Fine-tune the LLM

For capabilities that are:

- Stable
- Generalizable
- Useful on every run
- Better exercised by default

Examples: prefer stronger edits and trajectories

BAD | BARCELONA AI & DATA SYSTEMS

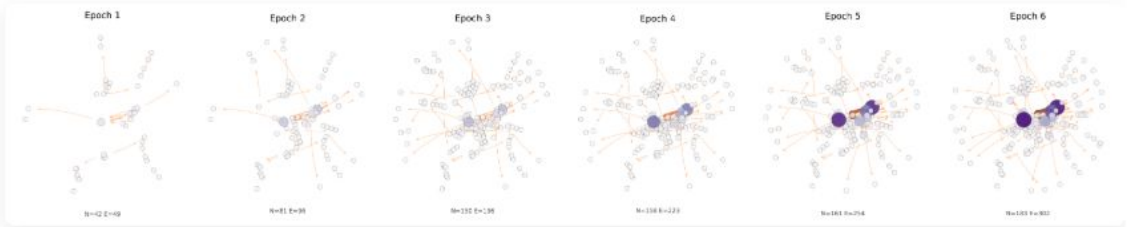
Evaluation: Continuous Learning from World Model

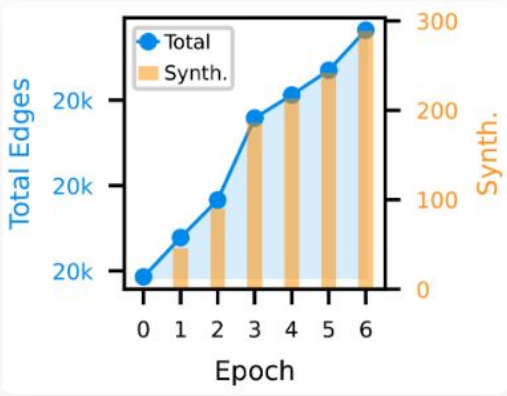
Adaptive traversal reinforcement, evaluated with 163 incidents across 6 epochs

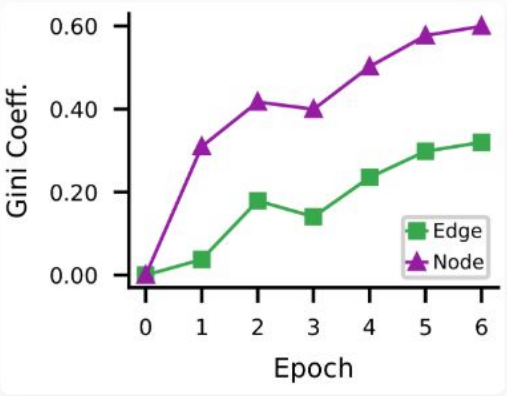
Table 5: Reinforcement Ablation (163 incidents, last 45 days)

Metric	With Reinf.	Without Reinf.
Reach Mitigation	**98.2%**	94.5%
Grounded	**92.6%**	89.0%
Avg. Usefulness	**3.45**	3.40
Avg. Similarity	**3.63**	3.51
Avg. Recall	**0.224**	0.207
Avg. Precision	0.217	**0.220**
KQL Success Rate	**93.8%**	89.2%
KQL Has-Rows Rate	**66.5%**	64.9%
Avg. KQL Succeeded	**12.6**	11.3

Graph Growth (6 epochs)
Nodes: 42 → **183** | Edges: 49 → **302**







BAD | BARCELONA AI & DATA SYSTEMS

From Darwin lineages to an improved coding policy

Objective software measurements become preference supervision.

1

Darwin lineage

Parent + sibling edits
+ objective scores

2

Preference data

Chosen edit vs.
rejected edit

3

Offline DPO

Qwen3.6 27B
+ LoRA adapter

4

Deploy in Darwin

Use tuned policy
as mutator agent

Current: offline preference learning **Next: Darwin rollouts + objective reward for on-policy learning**

BAD | BARCELONA AI & DATA SYSTEMS

Fine-tuning changes how the agent searches

The tuned policy reaches strong solutions with substantially less evolutionary search.

72.5%

fewer mutator-agent calls

to match the base policy's best quality on circle packing

Base policy

40

mutator agents

best valid score: 2.618

Tuned policy

11

mutator agents

best valid score: 2.631

Focused early result: 11 calls vs. 40 to match base-best quality; final score 2.631 vs. 2.618.

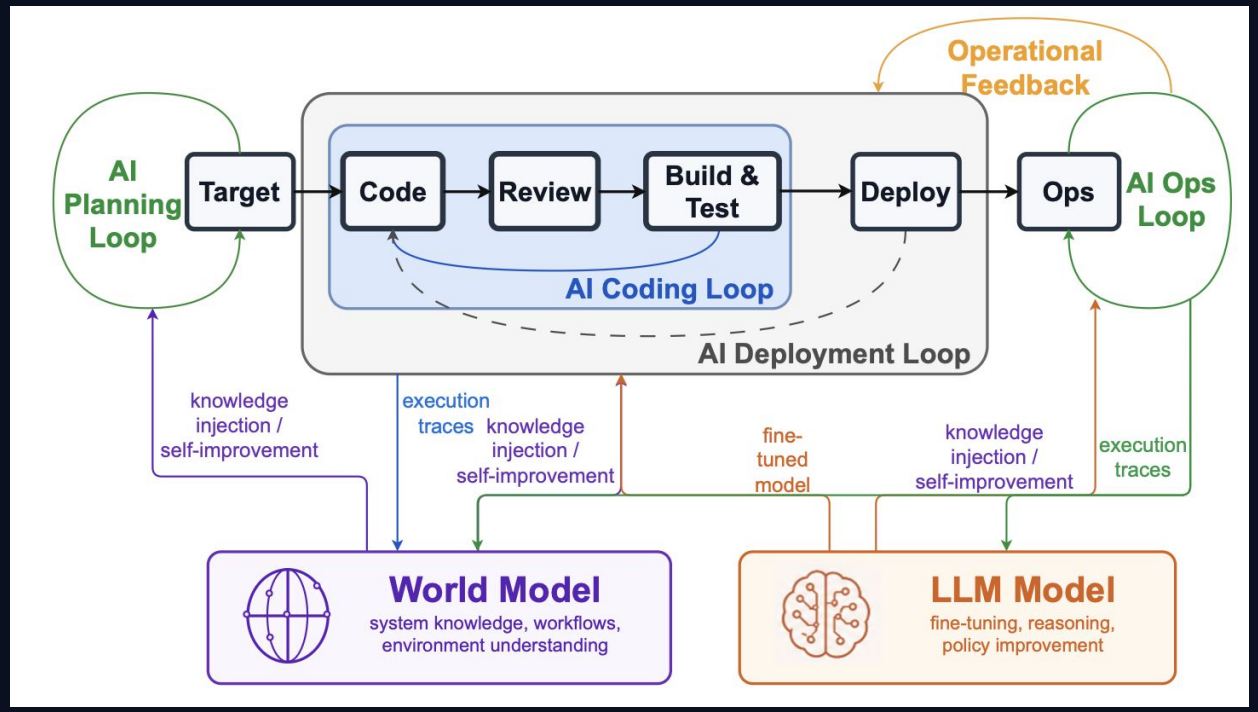
TAKEAWAY

BAD | BARCELONA AI & DATA SYSTEMS

The AI Software Factory improves software, operations, and itself

Optimize — Darwin

Target — Hotspan



Operate — AutoSRE

Learn — World Model + LLM

Thanks!

PANEL

BAD | BARCELONA AI & DATA SYSTEMS

AI Built in Barcelona, for the World

How can Barcelona turn its world-class research, entrepreneurial energy, and enterprise-scale engineering into AI products and systems with global impact?

Carlos Casanova (Glovo), **Laura Pereira Sánchez** (Microsoft), **Jesus Omaña Iglesias** (Telefónica)

BAD | BARCELONA AI & DATA SYSTEMS

THANK YOU! We hope the meetup was very… BAD!

See you in November at Universitat Politècnica de Catalunya (UPC) for the 3rd BAD Meetup!

Let us know any feedback and keep the conversation going in the BAD Discord!