ΚΟΜΒΙΚΗ ΤΕΤΑΡΤΗ

Open-Local-LLM

Τοπική υποδομή inference με ανοιχτό λογισμικό

και μοντέλα ανοιχτών βαρών για Δήμους, Περιφέρειες & Υπουργεία

Κοινή πρωτοβουλία

seatbase · Χίλων Πληροφορική · ΕΕΛΛΑΚ

Διαδικτυακή παρουσίαση | GR digiGOV-innoHUB | Μάιος 2026

Ποιοι είμαστε

Κοινή πρωτοβουλία τριών οργανισμών για ανοιχτή ψηφιακή υποδομή

seatbase

Hardware integration και sovereign infrastructure. Σχεδίαση ετερογενούς inference pool με Apple Silicon και NVIDIA nodes, κοινό API gateway, MLX/vLLM/SGLang serving και edge AI deployments.

seatbase.io

Χίλων Πληροφορική

Υλοποίηση και υποστήριξη παραγωγικών συστημάτων στον ελληνικό δημόσιο τομέα. Managed services, SLAs, integration με υφιστάμενα ΠΣ δήμων, υπουργείων και εποπτευόμενων φορέων.

hilonsys.com

ΕΕΛΛΑΚ

Οργανισμός Ανοιχτών Τεχνολογιών. Συντονισμός, θεσμική υποστήριξη ανοιχτής τεχνολογίας, ευρωπαϊκή δικτύωση, εκπαίδευση IT στελεχών δημοσίου, παρακολούθηση κανονιστικών εξελίξεων.

eellak.gr

Παρουσίαση: Στελέχη ΕΕΛΛΑΚ | Διάρκεια: 30 λεπτά (25 παρουσίαση + 5 ερωτήσεις)

Ατζέντα

ΜΕΡΟΣ 1

Το πρόβλημα

Σήμερα στην Ελλάδα:

>10 εκ. €

ενδεικτική υπόθεση κόστους για χρήση cloud LLM από δημόσιους φορείς

0%

έλεγχος πάνω στα δεδομένα μετά την αποστολή τους σε τρίτο πάροχο

~30%

ετήσια αύξηση τιμολόγησης ανά token τα τελευταία τρία χρόνια

. . .

Τέσσερις σύνθετοι κίνδυνοι

για κάθε δημόσιο φορέα που εξαρτάται από κλειστούς cloud LLM παρόχους

Vendor lock-in

Δυσχεραίνει τη μετάβαση σε εναλλακτικές. Όσο μεγαλώνει η εξάρτηση, τόσο μεγαλώνει το κόστος αλλαγής.

Απώλεια ελέγχου δεδομένων

Τα δεδομένα μπορεί να διασχίζουν εθνικά σύνορα και να υπόκεινται σε διαφορετικά νομικά καθεστώτα. Η διαγραφή, επιστροφή και χρήση τους εξαρτάται από τους όρους και τις συμβάσεις του παρόχου.

Αβεβαιότητα συμμόρφωσης

Οι όροι παροχής υπηρεσιών και οι τεχνικές πολιτικές μπορεί να αλλάζουν. Ο φορέας χρειάζεται συνεχή νομικό και τεχνικό έλεγχο για κάθε χρήση.

Διαρκής αύξηση κόστους

Τιμολόγηση ανά token κάνει αδύνατη τη μακροπρόθεσμη πρόβλεψη προϋπολογισμού. Όσο αυξάνεται η χρήση, αυξάνεται γραμμικά το κόστος.

Η εναλλακτική: Open-Local-LLM

Ψηφιακή κυριαρχία μέσω ανοιχτής, τοπικής, βιώσιμης υποδομής

Επένδυση σε τοπική υποδομή + υπηρεσίες λειτουργίας + μοντέλα ανοιχτών βαρών/αδειών.

Τα δεδομένα παραμένουν σε ελληνική επικράτεια ή ελεγχόμενη ευρωπαϊκή υποδομή.

01

Ψηφιακή κυριαρχία

Δεδομένα και μοντέλα σε εθνική υποδομή

02

Μείωση κόστους

Υποδομή ως επένδυση, όχι αυστηρά ως χρήση ανά token

03

Ενεργειακή απόδοση

Μέτρηση ανά workload και όχι γενική υπόσχεση

04

Υποστήριξη συμμόρφωσης

Auditability, traceability, human oversight

05

Εγχώρια τεχνογνωσία

Δεξιότητες παραμένουν στη χώρα

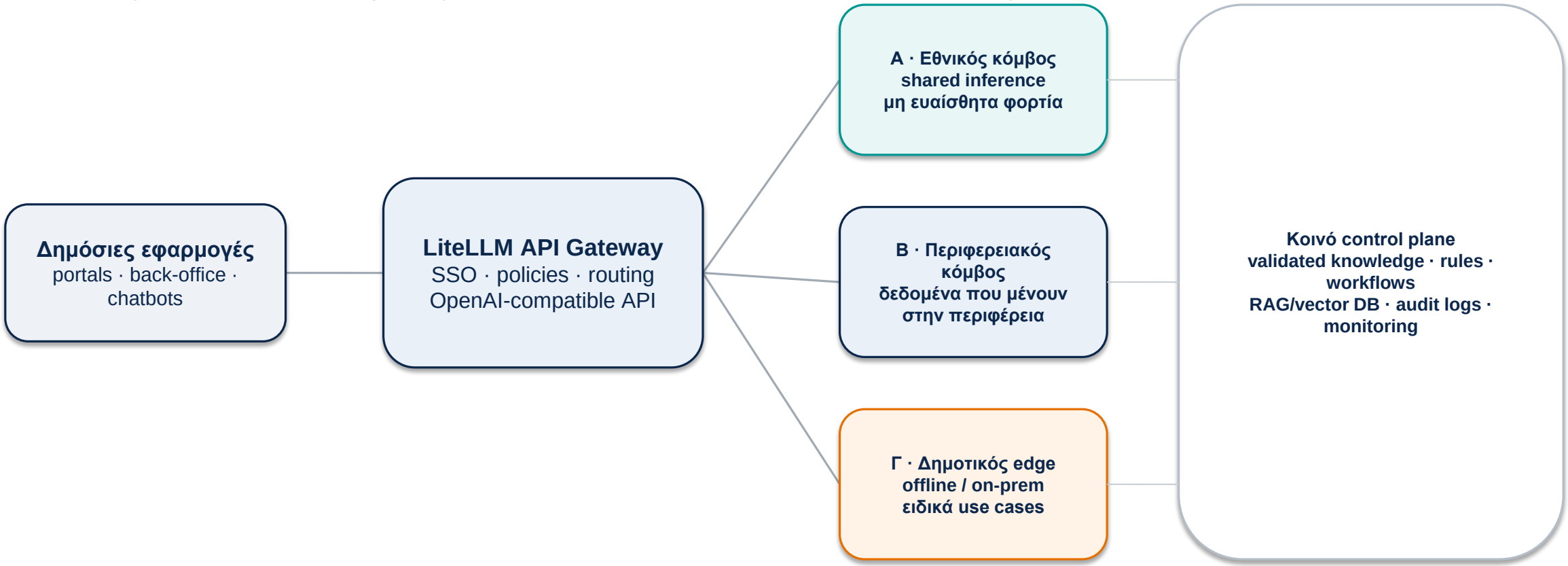
ΜΕΡΟΣ 2

Αρχιτεκτονική

Τρία επίπεδα ανάπτυξης

Αρχιτεκτονική σε μία εικόνα

Οι εφαρμογές μιλούν με ένα API· το gateway δρομολογεί ανάλογα με ευαισθησία δεδομένων, κόστος και φόρτο.



Κρίσιμο σημείο: Ενοποιούμε API, identity, logs, rules και πολιτικές. Για θεσμικά κρίσιμες διαδικασίες, το LLM βοηθά σε intake/επεξήγηση — δεν λειτουργεί ως decision engine.

Επίπεδο Α — Εθνικός Πάροχος LLM

Φιλοξενείται κεντρικά, λειτουργεί ως μοιρασμένη υπηρεσία inference

Τι είναι

Μία ετερογενής υποδομή inference από την ΕΕΛΛΑΚ / ΚΕΔΕ / ΕΝΠΕ, που παρέχει υπηρεσίες μέσω OpenAI-συμβατού API. Τα Apple Silicon και NVIDIA nodes συνυπάρχουν ως διαφορετικά backends πίσω από κοινό gateway. Η ενοποίηση γίνεται με routing, quotas, logs και monitoring — όχι με κοινή μνήμη GPU.

Σε ποιους απευθύνεται

- Δήμους <50.000 κατοίκων
- Υπουργεία για εσωτερικά εργαλεία
- Εποπτευόμενους φορείς, ερευνητικά κέντρα
- ΜμΕ μέσω πιστοποιημένου SaaS

Ενδεικτικό heterogeneous inference pool Επιπέδου Α

2–4×	**Mac Studio M3 Ultra 256/512GB**	*Large models*
2–4×	**NVIDIA L40S / RTX PRO 6000**	*high-throughput serving*
1×	**NVMe storage 30–50TB + backup**	models / RAG / logs
1×	**10/25/40 GbE networking**	API / storage / monitoring
—	**UPS, backup, monitoring, support**	operations

Συνολικό CAPEX (προς επικύρωση)

90 — 140 χιλ. €*

Πώς συνδυάζονται τα HW resources

Ετερογενές inference pool: routing architecture, όχι ενιαίο shared-memory/GPU cluster

1 · Κοινό API & routing

- OpenAI-compatible endpoint
- routing ανά φορέα, use case και κόστος
- quotas, rate limits, fallback
- model registry και policies

2 · Apple Silicon pool

- Mac Studio M3 Ultra / M4 Max
- MLX, llama.cpp, Ollama
- large quantized και offline workloads
- όχι κοινή GPU μνήμη με NVIDIA

3 · NVIDIA pool

- L40S / RTX PRO / server GPUs
- vLLM, SGLang, TGI
- continuous batching, PagedAttention
- RadixAttention και tensor parallelism

4 · Storage & ops

- NVMe models, RAG indexes, logs
- Qdrant, PostgreSQL, Redis
- Prometheus, Grafana, Loki
- backup, monitoring, support

Η ενοποίηση γίνεται στο επίπεδο gateway και observability: τα workloads δρομολογούνται στο κατάλληλο backend, δεν “ενώνονται” όλα τα resources σε ένα ενιαίο GPU cluster.

Επίπεδο Β — Περιφερειακός κόμβος LLM

Ένας ανά Περιφέρεια. Ευαίσθητα δεδομένα παραμένουν εντός περιφέρειας

13 Περιφέρειες

Κάθε περιφερειακός κόμβος εξυπηρετεί τους εντός περιφέρειας δήμους ως δευτερεύουσα μοιρασμένη υπηρεσία για ευαίσθητα δεδομένα και τοπικά RAG workloads.

Αττική	Κεντρ. Μακεδονία	Δυτ. Ελλάδα
Πελοπόννησος	Στ. Ελλάδα	Θεσσαλία
Ήπειρος	Δυτ. Μακεδονία	Αν. Μακεδονία—Θράκη
Κρήτη	Β. Αιγαίο	Ν. Αιγαίο
Ιόνια Νησιά		

Ενδεικτική υποδομή ανά Περιφέρεια

1–2×	**Mac Studio M4 Max 128GB ή M3 Ultra 96/256GB**	local LLM
1×	**NVIDIA L40S / RTX PRO class**	serving
—	**NVMe 12–24TB + δίκτυο**	RAG / logs
—	**UPS + monitoring + backup**	PoP-grade

Κόστος ανά Περιφέρεια (ενδεικτικό)

20 — 45 χιλ. €*

Το τελικό εύρος εξαρτάται από GPU, μνήμη, SLA και προμήθεια.

Επίπεδο Γ — Δημοτικός Edge κόμβος

Ένας υπολογιστής Apple Silicon σε γραφειακό περιβάλλον — όχι datacenter

Τυπική διαμόρφωση

Ένας Mac Studio M4 Max 64/128GB ή M3 Ultra 96/256GB σε γραφειακό περιβάλλον. Καμία ειδική ψύξη, κανένα rack, κανένα datacenter περιβάλλον.

- Φιλοξενεί 7B–14B μοντέλα άνετα και 30B-class μοντέλα με κβαντοποίηση, ανά benchmark
- Η κατανάλωση μετριέται ανά workload, όχι ως γενική υπόσχεση
- Λειτουργεί σε γραφειακό περιβάλλον 10–35°C
- Δεν απαιτεί datacenter rack ή ειδική ψύξη
- Πλήρως offline λειτουργία αν χρειαστεί

Πότε χρειάζεται Επίπεδο Γ

- On-vehicle inference (PotholeVision)
- CitizenVoice σε social media data
- Πιλοτικά εξ ολοκλήρου offline
- Δήμοι με αυστηρές εσωτερικές πολιτικές

Κόστος ανά κόμβο (ενδεικτικό)

4,7 — 12 χιλ. €*

Συμπληρώνει — δεν αντικαθιστά τα Επίπεδα A & B

Πηγή για ενδεικτική κατανάλωση/λειτουργία Mac Studio: Apple Support & Technical Specifications. Η inference κατανάλωση πρέπει να μετρηθεί στο pilot.

Sovereign Inference Node

Η θεμελιώδης δομική μονάδα του Open-Local-LLM

Apple Silicon — large-memory local inference

- MLX · llama.cpp · Ollama
- Large quantized models και offline/on-prem workloads
- 70B-class serving μόνο μετά από benchmark
- Χαμηλός θόρυβος και γραφειακή λειτουργία για edge nodes
- Δεν μοιράζεται GPU memory με NVIDIA nodes
- Ρόλος: ειδικά/τοπικά workloads, όχι production throughput για πολλά concurrent users

NVIDIA — production throughput

- vLLM · SGLang · TGI σε L40S / RTX PRO / server GPUs
- Continuous batching για πολλά ταυτόχρονα αιτήματα
- PagedAttention για αποδοτική διαχείριση KV cache
- RadixAttention / prefix caching για κοινά prompts και RAG templates
- Tensor parallelism μόνο σε ομοιογενές NVIDIA setup
- Ρόλος: shared APIs, high concurrency, predictable SLAs

Έξι τεχνικές βελτιστοποίησης

Ποιες τεχνικές μειώνουν τις απαιτήσεις υποδομής και αυξάνουν throughput

01	**4-bit / 8-bit quantization**	Μείωση μνήμης· η ποιότητα ελέγχεται ανά use case
02	**Continuous batching**	Υψηλότερο throughput όταν υπάρχουν πολλά αιτήματα
03	**PagedAttention**	Αποτελεσματική διαχείριση KV cache σε paged blocks
04	**RadixAttention / prefix caching**	KV cache reuse για κοινά prefixes: system prompts, RAG templates, few-shot παραδείγματα
05	**Speculative decoding**	Ταχύτερη παραγωγή σε κατάλληλα workloads
06	**Tensor parallelism**	Διαμοιρασμός βαρών σε ομοιογενές NVIDIA setup — όχι Apple+NVIDIA μαζί

Λογισμικό στοίβα — production-grade

Ανοιχτό λογισμικό υποδομής + έλεγχος αδειών μοντέλων πριν από παραγωγή

Διεπαφή χρήστη

Open WebUI · LangChain · Haystack

Gateway & Ασφάλεια

LiteLLM · Keycloak · gov.gr SSO

Inference Engines

vLLM · SGLang/RadixAttention · TGI · MLX · llama.cpp · Ollama

Δεδομένα & RAG

Qdrant · ChromaDB · PostgreSQL · Redis

Παρακολούθηση

Prometheus · Grafana · Loki · OpenTelemetry

Ανάπτυξη

Docker · Kubernetes/k3s · Ansible · Gitea

Review αδειών ανά εργαλείο και μοντέλο: Apache-2.0/MIT/AGPL ≠ custom open-weight licenses. Production χρήση μόνο μετά από benchmark και license review.

Λειτουργική αρχιτεκτονική SW

Από εφαρμογή σε απάντηση: οι στρώσεις που χρειάζεται ένας παραγωγικός κόμβος

Εφαρμογές	portals · back-office · chatbots · batch jobs
Access & ασφάλεια	Keycloak / gov.gr SSO · RBAC · tenant policies · audit IDs
Gateway / router	LiteLLM · OpenAI-compatible API · quotas · fallback · cost tracking
Serving backends	NVIDIA: vLLM/SGLang/TGI | Apple: MLX/llama.cpp/Ollama
Knowledge, Rules as Code & RAG	Versioned rules/DMN · Qdrant/Chroma · PostgreSQL · citation store · document loaders
Ops & governance	Prometheus/Grafana/Loki · model cards · license registry · benchmarks · decision logs

Λογική επιλογής backend και ροής

- Θεσμικά κρίσιμη ροή → Rules as Code / deterministic workflow + audit trail
- Αναζήτηση/τεκμηρίωση → RAG + citations
- Υψηλό concurrency / shared APIs → NVIDIA
- Repeated templates / RAG prompts → SGLang

Κοινό API — θεσμικές αποφάσεις με Rules as Code, LLM μόνο για υποβοήθηση.

Μοντέλα LLM ανοιχτών βαρών

Model-agnostic αρχιτεκτονική — επιλογή μετά από benchmarks, license review και αξιολόγηση σε ελληνικά δεδομένα

70B+	**Νομικά / διοικητικά κείμενα** Llama 3.3 70B · Mistral Large 2 · Qwen 2.5 72B	**Mid-size**	**Document understanding, RAG** Qwen 2.5/3 32B · Mixtral 8×7B · QwQ 32B
8B–14B	**Γενικής χρήσης chatbots** Qwen 2.5/3 8B–14B · Aya Expanse · Gemma 2	**3B–8B**	**Classification / assistants** Qwen 2.5/3 3B–8B · Phi-4 Mini · Mistral 7B
GR	**Greek public-sector fine-tunes** Qwen 2.5/3 fine-tuned σε ελληνικά διοικητικά corpora · Krikri/Meltemi ως baselines	**ASR**	**Speech-to-text & embeddings** Whisper Large v3 · multilingual-e5 · BGE-M3 · Greek-BERT

RAG για ανάκτηση πηγών · fine-tuning για ύφος/εργασίες/structured output · θεσμική γνώση μόνο με versioned rules και expert validation.

Validated Knowledge Layer

LLM ως interface, όχι ως μηχανισμός απόφασης

1 · Versioned γνώση

- νόμοι, εγκύκλιοι, διαδικασίες
- ημερομηνίες ισχύος
- domain/legal expert validation
- change control

2 · Rules as Code

- BPMN/DMN workflows
- decision tables / rule engine
- eligibility checks
- rule ID + version ανά βήμα

3 · RAG / citations

- ανάκτηση σχετικών πηγών
- τεκμηρίωση & παραπομπές
- semantic search
- όχι authoritative decision

4 · LLM interface

- intake και ερωτήσεις χρήστη
- σύνοψη / draft απάντησης
- επεξήγηση σε απλή γλώσσα
- human review όπου απαιτείται

Validated Knowledge & Rules as Code Layer

Εκτελέσιμοι κανόνες και audit trail· το LLM δεν είναι decision engine

Η αρχιτεκτονική σε δράση: 30 πιλοτικά έργα ΤΝ

Συγκεκριμένα έργα για δήμους και περιφέρειες — με σαφή χρονικό ορίζοντα

30
πιλοτικά έργα ΤΝ
συνολικά

9
Rules as Code + LLM
για διοικητικά use cases

7
ελαφρύ NLP
(3B–14B)

14
CV / time-series / opt.
(χωρίς LLM)

Σειρά προτεραιότητας υλοποίησης

3 μήνες	6 πιλοτικά	Άμεση εκκίνηση — 100% ανοιχτά δεδομένα, χωρίς MOU
6 μήνες	13 πιλοτικά	Μεσοπρόθεσμα — απαιτούν 1-2 MOUs ή labeled data
12 μήνες	11 πιλοτικά	Ετήσιο πρόγραμμα — DPIA, IoT, εθνικές πρωτοβουλίες

Πιλοτικά με ασφαλή χρήση AI στη διοίκηση

Τα LLMs υποστηρίζουν intake, αναζήτηση, σύνοψη και επεξήγηση. Θεσμικές αποφάσεις/πιστοποιητικά εκτελούνται από Rules as Code και workflows.

Δ5 — **ePermits Assistant** — Rules

Eligibility checks με κανόνες · RAG μόνο για πηγές · LLM για επεξήγηση φακέλου

Δ6 — **BizLicense Match** — Rules

Αντιστοίχιση διαδικασιών με decision tables · επίσημες πηγές · καθοδήγηση χρήστη

Δ7 — **SmartTriage 155** — NLP

Whisper + classification · routing σε υπηρεσία · escalation σε υπάλληλο

Δ10 — **HR Assistant** — Rules

Decision tables για συχνές HR διαδικασίες · citations · human review

Δ11 — **Registry Workflow Assistant** — Workflow

Έκδοση από υφιστάμενο ΠΣ/ruleset · LLM μόνο για intake, πληρότητα, επεξήγηση

Δ13 — **SocialCare Radar** — Assistive

Anomaly signals + επεξήγηση για υπάλληλο · όχι αυτόματη απόφαση

Π8 — **VocationalMatch** — Recommend

Explainable matching κριτηρίων · recommendation assistant · όχι διοικητική απόφαση

Π9 — **Grant Screening Assistant** — Screening

Eligibility screening με κανόνες · LLM για σύνοψη φακέλων · όχι τελική κατανομή

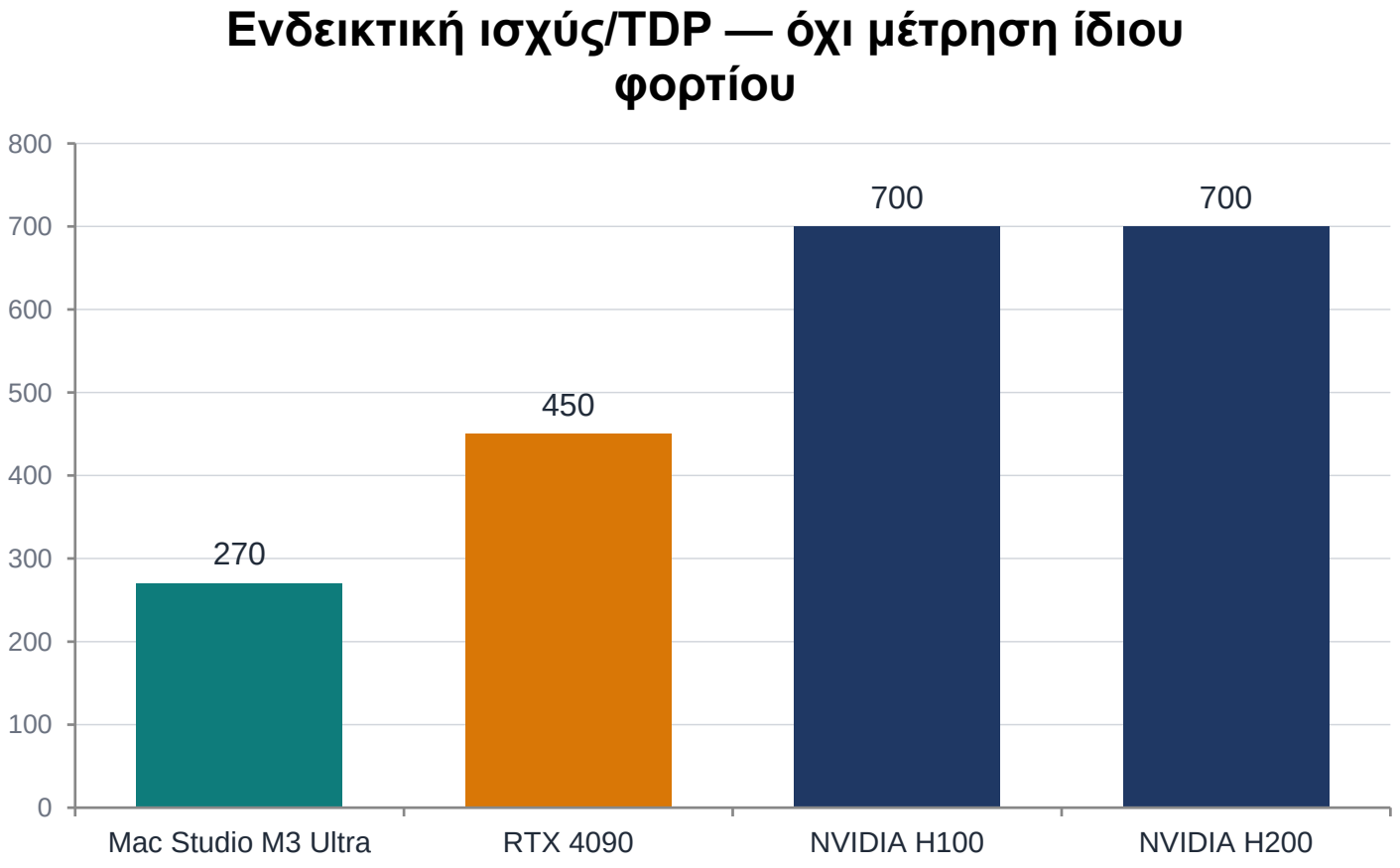
Π12 — **HealthEquity Map** — GIS

GIS/analytics + LLM επεξηγήσεις δεικτών · policy support μόνο

Όπου υπάρχει έννομη συνέπεια, κρατάμε traceable workflow: input → rule ID/version → νομική βάση → έλεγχος → αποτέλεσμα → human approval όπου απαιτείται.

Ενεργειακό αποτύπωμα

Πράσινες δημόσιες προμήθειες



Πράσινες προμήθειες

Μετρήσιμο

ενεργειακό όφελος

με ίδιο benchmark πριν από παραγωγή

- Δεν απαιτεί απαραίτητα datacenter cooling για edge κόμβους
- Συμβατότητα με GPP να τεκμηριωθεί ανά προμήθεια
- Μετρήσιμο ESG benefit

Σημείωση: οι τιμές του γραφήματος είναι ενδεικτικά όρια ισχύος/TDP από τεχνικά specs, όχι συγκριτικό benchmark inference. Το “3–5×” να παραμείνει στόχος προς μέτρηση, όχι δεδομένο.

Σχεδιασμός για συμμόρφωση

GDPR, EU AI Act, εθνική νομοθεσία -- η αρχιτεκτονική βοηθά, αλλά απαιτείται έλεγχος ανά εφαρμογή

GDPR — Καν. 2016/679

- Τοπική επεξεργασία — μειώνεται η ανάγκη εξαγωγής δεδομένων σε τρίτους παρόχους
- Ευαίσθητα δεδομένα Άρθρου 9 → Επίπεδο Β μόνο
- Δικαιώματα υποκειμένων μέσω audit logging
- Retention policy logs ανά φορέα και σκοπό επεξεργασίας
- DPIA όπου απαιτείται πριν από παραγωγική χρήση

EU AI Act — Καν. 2024/1689

- Model Cards για κάθε εφαρμογή
- Datasheets για κάθε μοντέλο σε χρήση
- Audit logging για record keeping
- Human oversight και Rules as Code / deterministic workflows για κρίσιμες αποφάσεις
- Περιοδικός έλεγχος μεροληψίας για high-risk use cases
- Διαφάνεια χρήσης ΤΝ προς πολίτες (Άρθρο 50)

Η τοπική υποδομή υποστηρίζει τη συμμόρφωση, αλλά δεν την “εξασφαλίζει” μόνη της. Σε θεσμικές διαδικασίες χρειάζονται DPIA, risk assessment, Rules as Code / versioned rules, audit trail και ανθρώπινη εποπτεία.

Φάσεις υλοποίησης

Από πιλοτική εφαρμογή σε εθνική κυρίαρχη υποδομή ΤΝ

ΦΑΣΗ 1

Μήνες 1-6

Πιλοτική εφαρμογή

- 3 Πανεπιστήμια
- 2 Περιφέρειες
- 1 Υπουργείο
- 1 Δίκτυο ΜμΕ

ΦΑΣΗ 2

Μήνες 6-18

Ελληνικό LLM Αποθετήριο

- Ελληνικό Knowledge & Model Repository

ΦΑΣΗ 3

Μήνες 12-36

Δίκτυο Περιφερειακών AI Nodes

- Πανεπιστημιακοί κόμβοι
- Τεχνολογικά πάρκα
- Δημοτικά data centers
- ΑΠΕ campuses

Πώς συμμετέχετε

Συγκεκριμένα βήματα για δήμους, περιφέρειες και υπουργεία

1

Διαβάστε την κοινή τεχνική πρόταση

32-σέλιδο έγγραφο seatbase · Χίλων · ΕΕΛΛΑΚ — διαθέσιμο σε όλους τους ενδιαφερόμενους ως DRAFT για σχολιασμό

2

Συμπληρώστε τα 21 ερωτήματα του εγγράφου

Αυτο-αξιολόγηση: ποιο επίπεδο σας ταιριάζει, ποιες εφαρμογές προτεραιοποιείτε, ποιοι περιορισμοί υπάρχουν

3

Ορίστε IT champion στον οργανισμό σας

Ένα στέλεχος που γνωρίζει το tech landscape και μπορεί να σχεδιάσει 1-2 πιλοτικές εφαρμογές

4

Συντονιστείτε με τις Φάσεις 1-3

Εκδήλωση ενδιαφέροντος για Φάση 1 ως πιλοτικός φορέας ή για ένταξη σε επόμενες φάσεις

ΣΥΝΟΨΗ

Τι να θυμάστε

1. Data + models δεν αρκούν: χρειάζεται επικυρωμένη γνώση, κανόνες, workflows, άδειες και benchmarks.
2. Με εκτιμώμενο αρχικό CAPEX, μπορεί να ξεκινήσει κλιμακούμενη εθνική υποδομή inference για δημόσιους φορείς.
3. Τα resources συνδυάζονται ως ετερογενές inference pool πίσω από κοινό API — όχι ως ένα ενιαίο GPU cluster.
4. Σε θεσμικά κρίσιμες ροές, οι αποφάσεις παράγονται από Rules as Code / workflows — το LLM υποστηρίζει μόνο εξήγηση και τεκμηρίωση.
5. Η μετάβαση μειώνει εξάρτηση από τρίτες χώρες, δημιουργεί εγχώρια προστιθέμενη αξία και διατηρεί τεχνογνωσία στη χώρα.

Ευχαριστούμε

Ερωτήσεις & συζήτηση

seatbase

Hardware integration
Sovereign infrastructure

seatbase.io

Χίλων Πληροφορική

Παραγωγική υλοποίηση
Managed services

hilonsys.com

ΕΕΛΛΑΚ

Συντονισμός, Θεσμική υποστήριξη
Εκπαίδευση IT στελεχών

eellak.gr

Παρουσιάστηκε στο πλαίσιο της Κομβικής Τετάρτης 20/5/2026 του GR digiGOV-innoHUB

Έγγραφο τεχνικής τεκμηρίωσης (32 σελίδες, DRAFT v1.0) διατίθεται για σχολιασμό