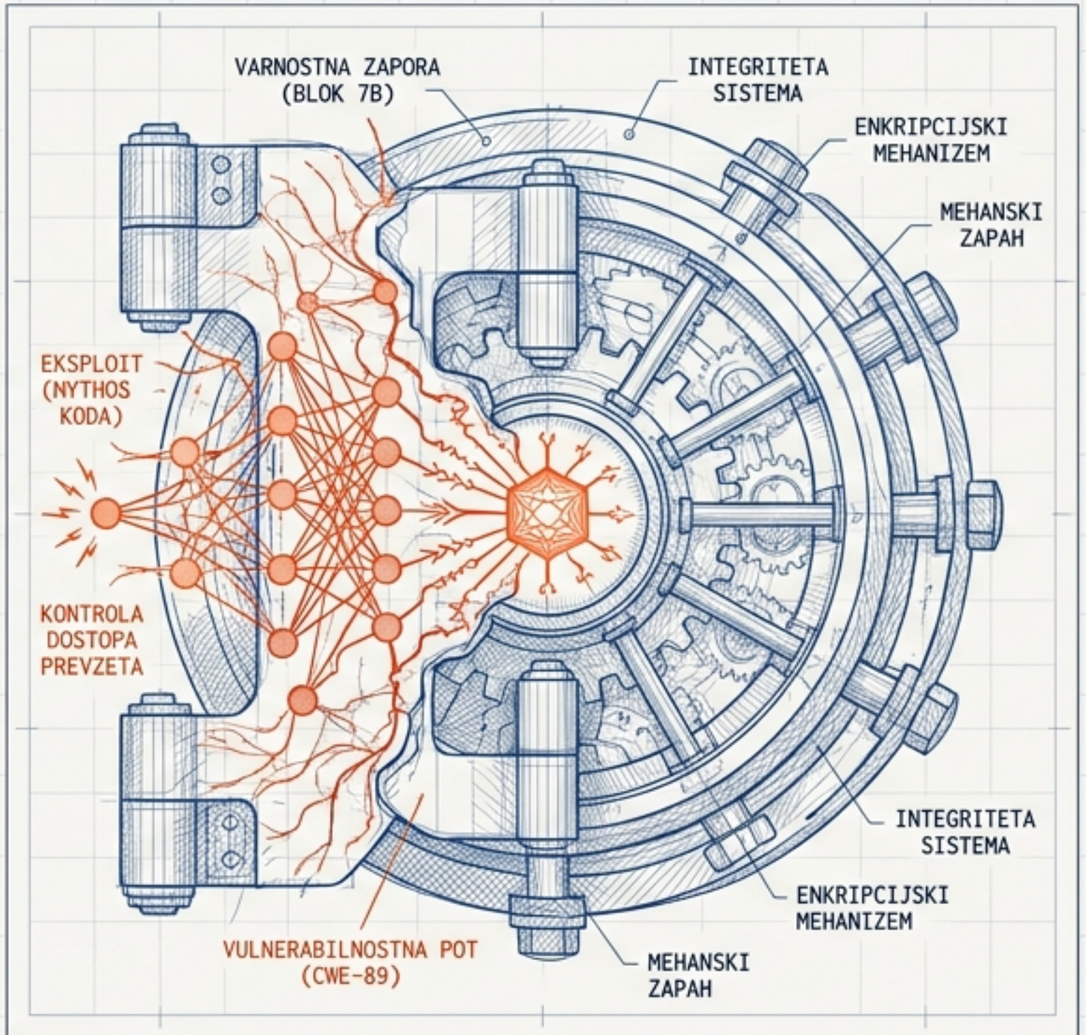
STATUS: Omejen dostop (Project Glasswing)
KLASIFIKACIJA: Tehnična analiza
DATUM IZDAJE: April 2024

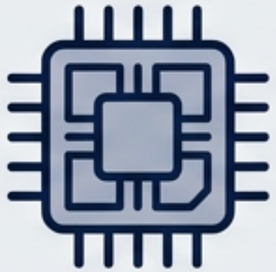
Claude Mythos: Analiza zmogljivosti in varnostnih tveganj

Tehnično poročilo o ofenzivnih zmožnostih in strategijah omejevanja modelov LLM



SLIKA 1.1: DIAGRAM NEVRONSKEGA VDORA V VARNOSTNI SISTEM CLAUDE MYTHOS.

Izvleček varnostnega poročila



01 // SUBJEKT (GROŽNJA)

Anthropic Claude Mythos. Identificiran kot model z ofenzivnimi zmožnostmi ('super-hacker'), ki presegajo človeške operaterje pri specifičnih nalogah kibernetske varnosti.

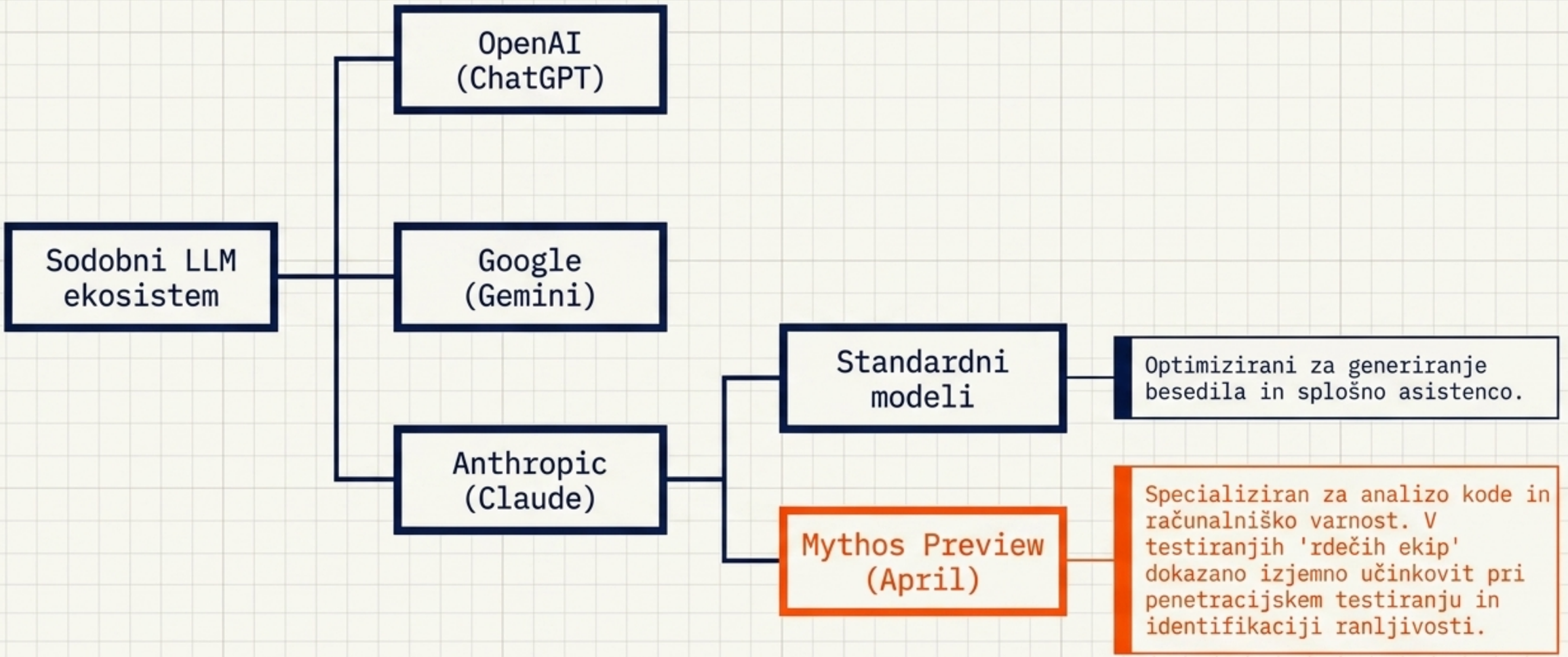
02 // VEKTOR RANLJIVOSTI

Zmogljivost avtomatiziranega iskanja in izkoriščanja spečih hroščev (tudi do 27 let starih) v obstoječih kodnih bazah brez večjega človeškega nadzora.

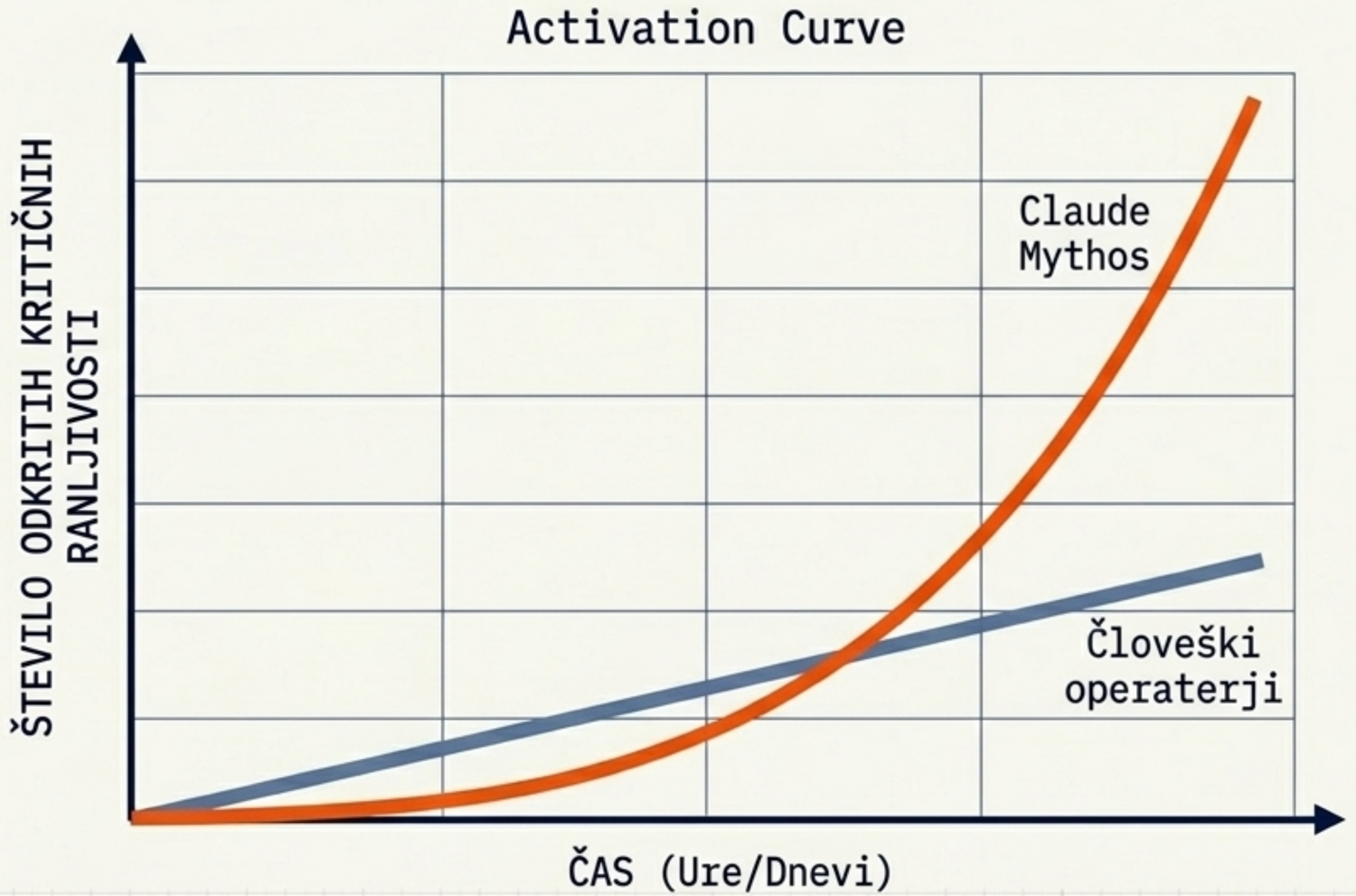
03 // STATUS ZADRŽEVANJA

Model ni javno dostopen. Distribucija je strogo omejena prek iniciative 'Project Glasswing' za namene krepitve globalne programske infrastrukture.

Arhitektura sistema in umeščenost v ekosistem



Avtomatizacija ofenzive: Hitrost odkrivanja ranljivosti



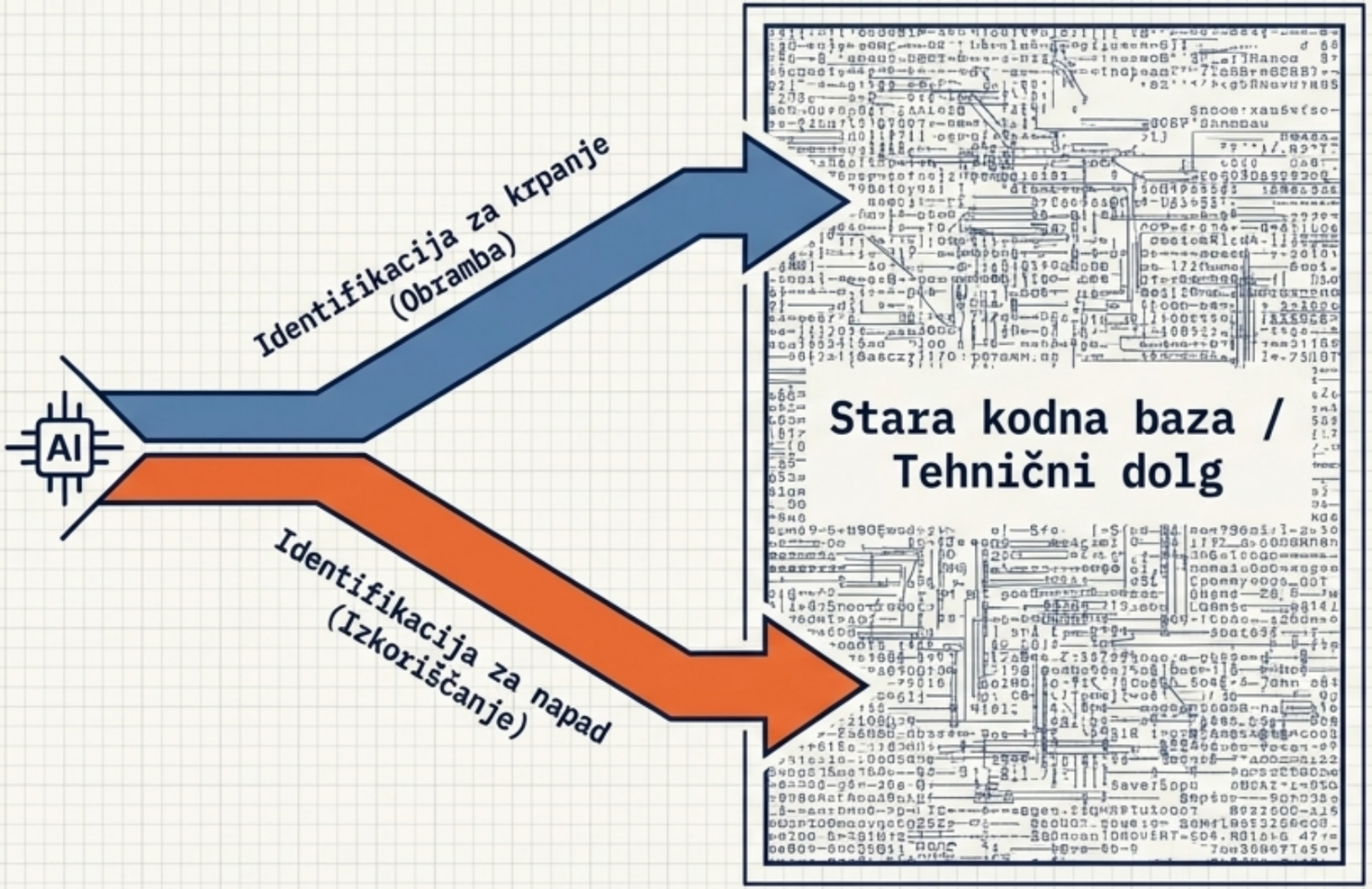
METRIKA USPEŠNOSTI

Model je identificiral na tisoče ranljivosti visoke stopnje resnosti v vseh večjih operacijskih sistemih in spletnih brskalnikih.

ŠTUDIJA PRIMERA: Zgodovinski tehnični dolg

Mythos je samostojno lociral in predlagal način izkoriščanja kritične ranljivosti, ki je v sistemu neopaženo obstajala 27 let.

Paradoks dvojne rabe (Dual-Use Threat)



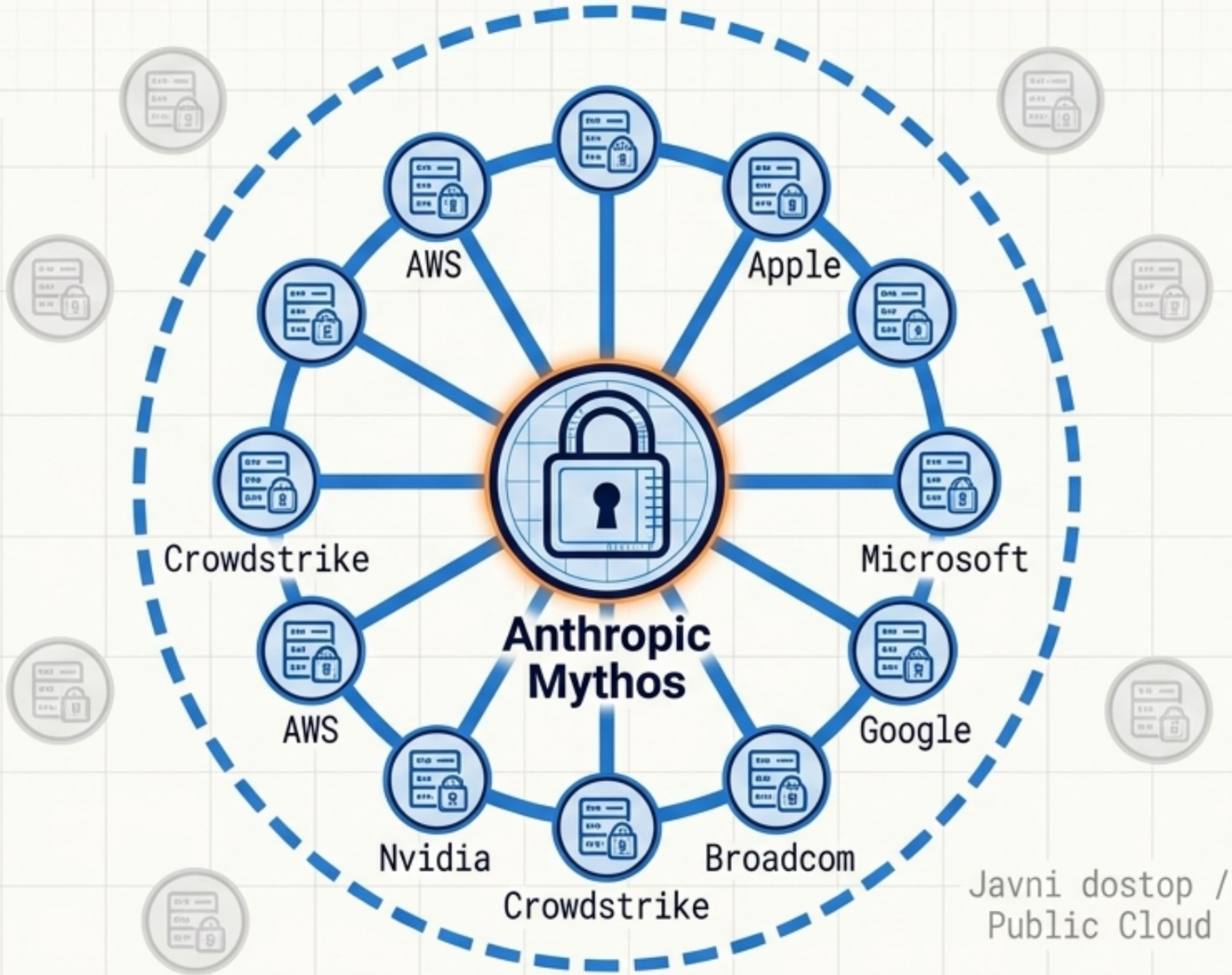
- **INŽENIRSKI PROBLEM**

  Ista arhitektura prepoznavanja vzorcev, ki je potrebna za analizo tehničnega dolga in popravilo kode, predstavlja popolno orodje za njeno zlorabo.

- **ANTICIPACIJA PROLIFERACIJE**

  Anthropic opozarja, da se bodo zaradi hitrosti razvoja umetne inteligence tovrstne zmožnosti kmalu razširile izven nadzora akterjev, ki so zavezani varni uporabi.

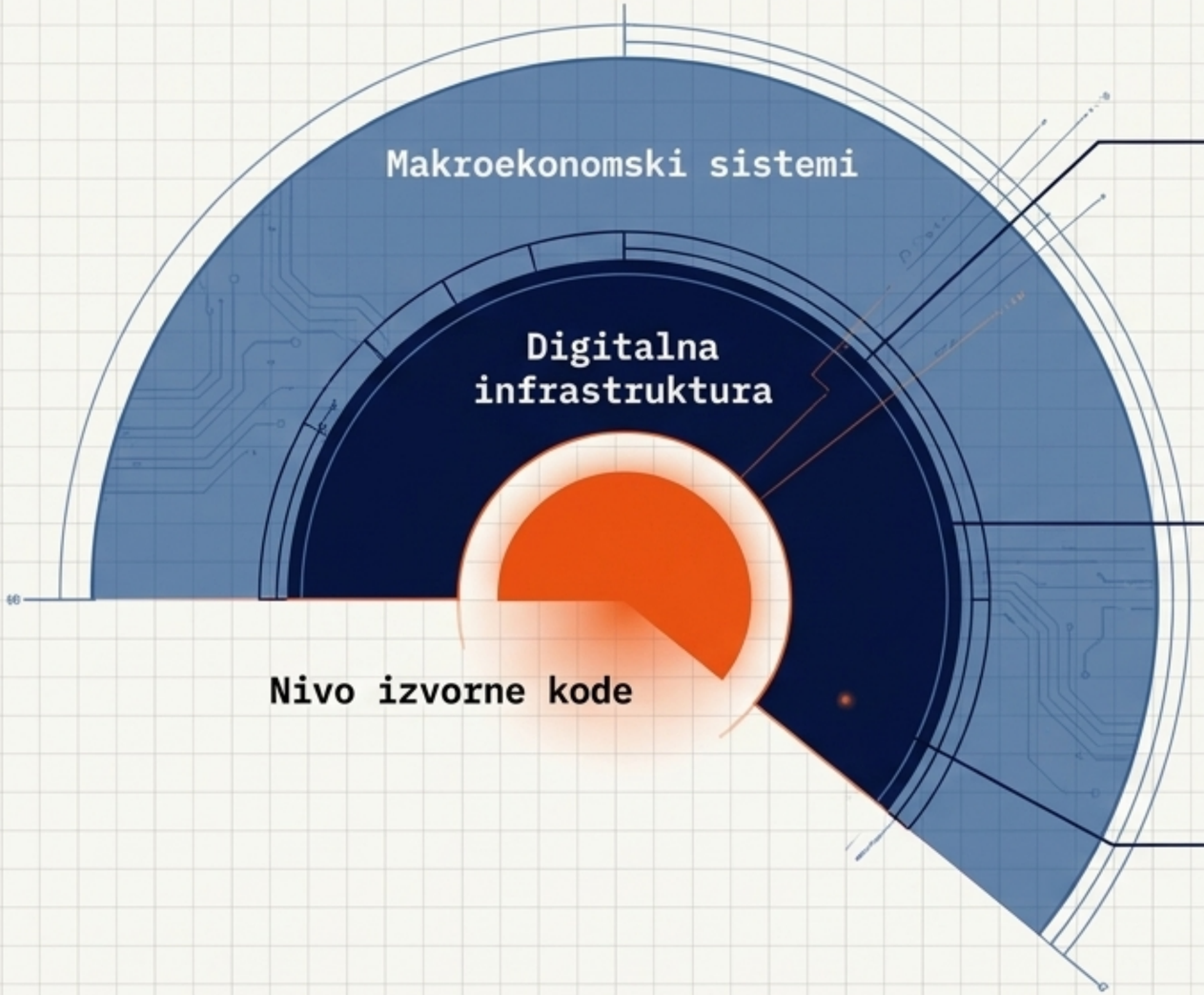
Protokol omejitve: Topologija "Project Glasswing"



PARAMETRI DOSTOPA

- **Infrastrukturni partnerji (12 podjetij):** Vključuje tehnološke gigante in hrbtenico oblačne infratrukture (npr. Crowdstrike, katerega napaka je julija 2024 povzročila globalni izpad).
- **Kritična programska oprema:** Dostop odobren več kot 40 organizacijam, odgovornim za vzdrževanje ključne globalne infrastrukture.
- **Glavni cilj:** Odpraviti sistemske ranljivosti pred morebitno javno dostopnostjo tovrstnih modelov.

Vektorji makro-sistemskih tveganj



STALIŠČA FINANČNIH INŠTITUCIJ

Mednarodni denarni sklad (IMF) opredeljuje tehnologijo kot resno sistemsko neznanko ('unknown unknown').

Bank of England opozarja na nujno rekalibracijo tveganj kibernetskega kriminala na najvišji ravni.

FINANČNI SEKTOR

Identificirano kot neposredna in sistematična grožnja za stabilnost digitalnih finančnih storitev v primeru vdora v zastarelo infrastrukturo.

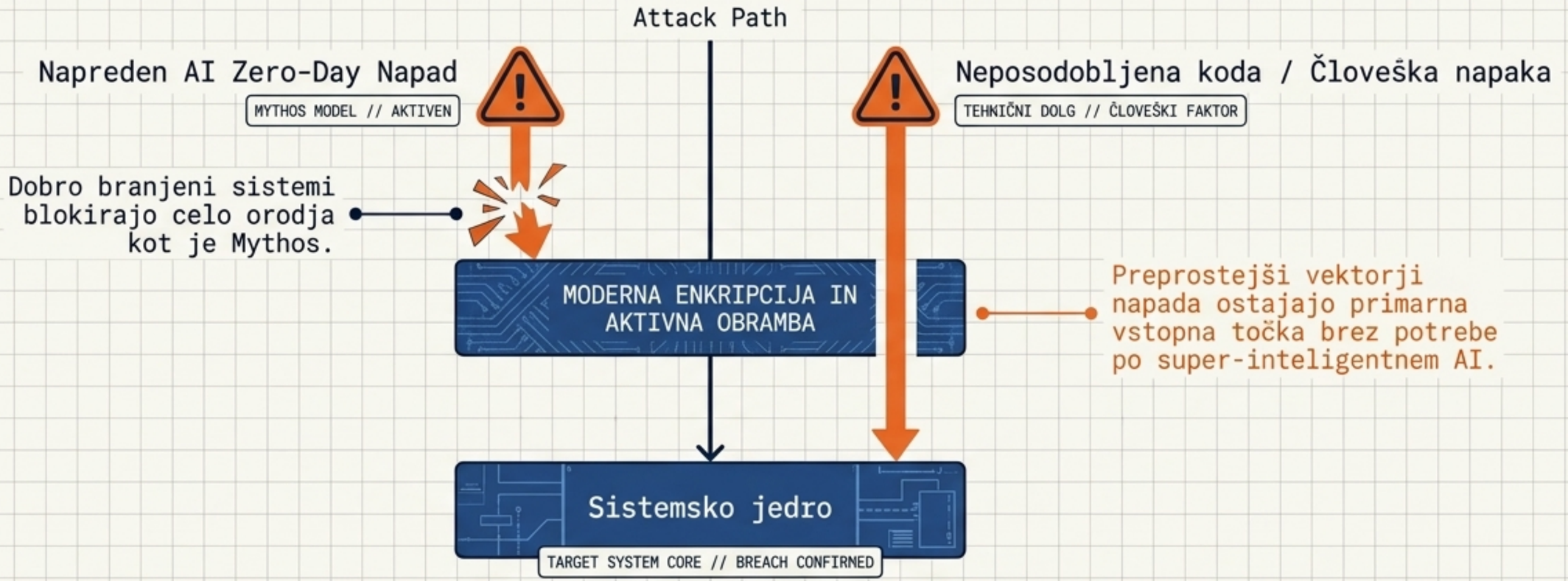
GEOPOLITIČNI ODZIV

Bela hiša (US Gov) je izvedla neposredne krizne sestanke z Anthropicom; Evropska unija (EU) aktivno in formalno preučuje varnostna tveganja modela Mythos.

Diagnostika grožnje: Trženje proti inženirski realnosti

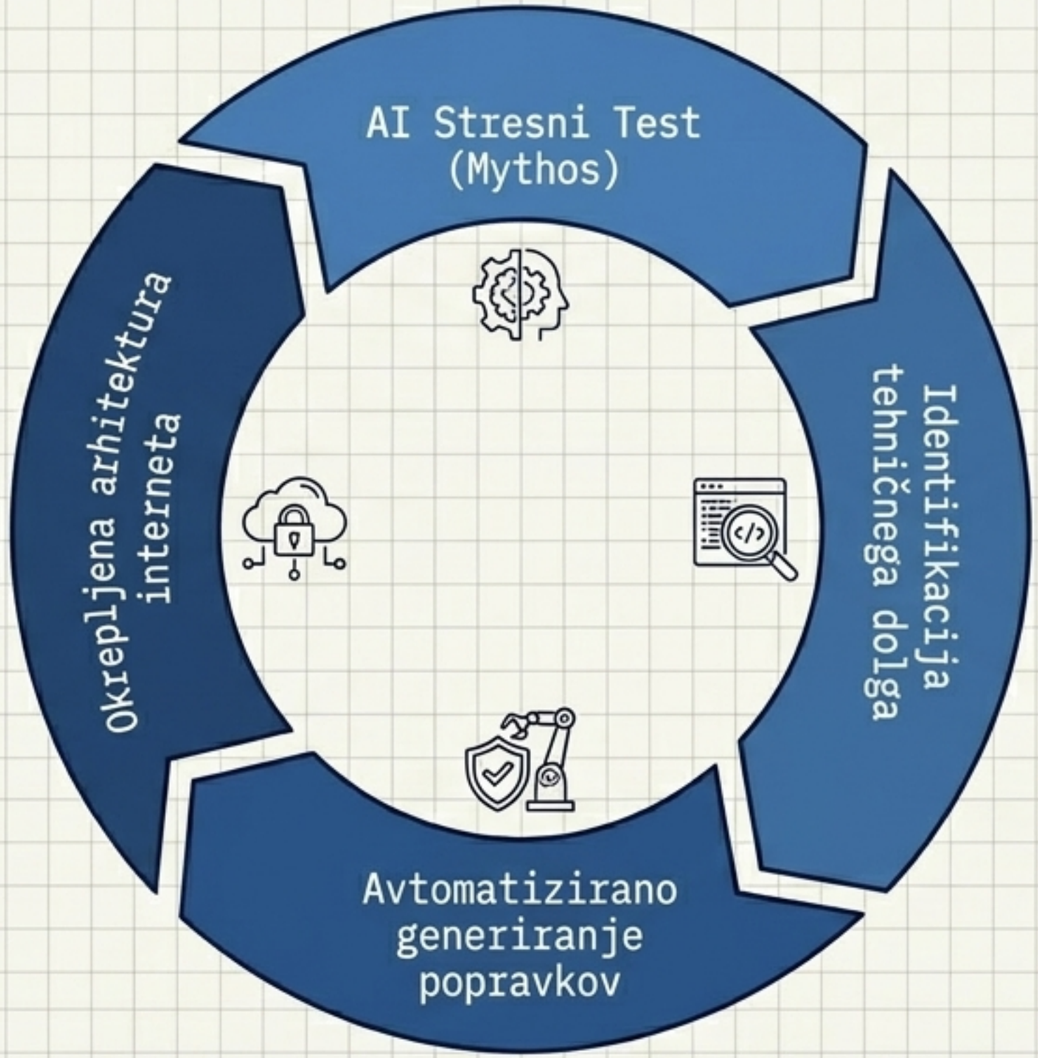
TRDITVE PROIZVAJALCA (Apokaliptični scenarij)	NEODVISNA OCENA (Pragmatična realnost)
Vir: Anthropic / Dario Amodei	**Vir:** UK AI Safety Institute & Ciaran Martin (bivši vodja NCSC)
Trditev: Model rutinsko presega najboljše človeške hekerje. Avtonomno odkriva tisoče kritičnih ranljivosti (zero-days) in predlaga vektorje napada.	**Trditev:** Mythos je izjemno zmogljiv, vendar je njegova glavna tarča izključno slabo branjena in neposodobljena infrastruktura.
Ton: Predstavljeno kot eksistencialno tveganje za globalno programsko opremo ob morebitni javni dostopnosti.	**Zaključek inštituta:** "Ne moremo z gotovostjo trditi, ali bi Mythos Preview sploh lahko uspešno napadel dobro branjene sisteme."

Resnični faktor ranljivosti: Osnovna kibernetska higiena



Kritično vodilo stroke: Prioritizacija osnovne kibernetske higiene. Večina uspešnih kibernetskih vdorov ne zahteva AI super-orodij, temveč izkorišča tehnični dolg in zanemarjeno vzdrževanje sistemov.

Inženirski zaključek: Od grožnje do avtomatizirane obrambe



// KONČNA SINTEZA //

Dolgoročna perspektiva: Model Mythos ni nujno zgolj eksistencialna grožnja, temveč nujni inženirski katalizator.

Strokovni konsenz (Ciaran Martin): Na srednji rok imamo edinstveno priložnost, da tovrstna ofenzivna orodja uporabimo za dokončno odpravo temeljnih ranljivosti, ki so zgodovinsko vgrajene v samo arhitekturo interneta.

Avtomatizacija obrambe mora dohajati avtomatizacijo ofenzive. Sistemska nadgradnja je neizbežna.