ANDORRA DIGITAL

ANDORRA DIGITAL
CIBERSEGURETAT

Informe de Ciberintel·ligència

Claude Mithos: anàlisi de les amenaces i les capacitats

TLP: WHITE

MAIG 2026

FITXA DEL DOCUMENT

Versió	Redactat/Revisat per	Aprovat per	Data aprovació	Data publicació
1.0	ANC-AD	ANC-AD	29/05/2026	03/06/2026

Registre de canvis			
Versió	Pàgines	Data Modificació	Motiu del canvi

Propietari del document	ANC-AD

ÍNDEX

1. METODOLOGIA

Aquest informe aplica els principis de Traffic Light Protocol (TLP). És un esquema creat per fomentar un intercanvi més bo d'informació delicada (però no classificada) en l'àmbit de la seguretat de la informació.

A través d'aquest esquema, d'una manera àgil i senzilla, s'indica fins on pot circular la informació més enllà del receptor immediat, i aquest ha de consultar l'Agència Nacional de Ciberseguretat d'Andorra quan cal distribuir la informació a tercers.

Codi	Com es fa servir	Com es comparteix
TLP: RED	S'ha de fer servir **TLP:RED** quan la informació està limitada a persones concretes, i podria tenir impacte en la privacitat, la reputació o les operacions si es fa servir malament.	Els receptors no han de compartir informació designada com a **TLP:RED** amb cap tercer fora de l'àmbit on va ser exposada originalment.
TLP: AMBER	S'ha de fer servir **TLP:AMBER** quan la informació ha de ser distribuïda de manera limitada, però suposa un risc per a la privacitat, la reputació o les operacions si és compartida fora de l'organització.	Els receptors poden compartir informació indicada com a **TLP:AMBER** només amb membres de la seva pròpia organització que necessiten conèixer-la, i amb clients, proveïdors o associats que necessiten conèixer-la per protegir-se a si mateixos o evitar danys. L'emissor pot especificar restriccions addicionals per compartir aquesta informació.
TLP: GREEN	S'ha de fer servir **TLP:GREEN** quan la informació és útil per a totes les organitzacions que hi participen, com també amb tercers de la comunitat o el sector.	Els receptors poden compartir la informació indicada com a **TLP:GREEN** amb organitzacions afiliades o membres del mateix sector, però mai a través de canals públics.
TLP: WHITE	S'ha de fer servir **TLP:WHITE** quan la informació no suposa cap risc de mal ús, conforme a les regles i procediments establerts per a la seva difusió pública.	La informació **TLP:WHITE** pot ser distribuïda sense restriccions, únicament subjecta a controls de copyright.

2. INTRODUCCIÓ

El dia 7 d'abril del 2026, Anthropic, l'empresa nord-americana dedicada a la recerca i el desenvolupament de la intel·ligència artificial va publicar informació sobre Claude Mythos, el seu model d'IA més avançat fins ara. **El que diferencia el llançament de Claude Mythos no és únicament allò que el model és capaç de fer, sinó allò que ha fet sense que ningú li ho hagués demanat abans**.

Durant unes proves internes de l'empresa Anthropic, una versió primerenca de Claude Mythos **va escapar d'un entorn controlat, va obtenir accés a Internet i va enviar un mail a l'investigador que supervisava les proves**, que en aquell moment no estava supervisant res i no havia demanat res del que el model estava fent.

El següent informe té com a objectiu analitzar aquest incident i tot allò que envolta Claude Mythos, les seves capacitats tècniques, els riscos d'alineació que l'empresa Anthropic ha documentat del model, i tot allò que implica des d'una perspectiva de ciberseguretat.

Tot seguit, mostrarem una taula sobre què cobrirà l'informe i què no:

Què cobreix aquest informe
Les capacitats tècniques documentades de Claude Mythos Preview. *Incloent-hi* benchmarks*, descobriment de vulnerabilitats i l'incident de contenció.*
El model d'amenaça i els vectors de risc identificats per Anthropic. *Els sis* pathways *de risc prioritzats a l'Alignment Risk Update.*
Les mitigacions existents i les limitacions conegudes. *Incloent el Project Glasswing i els controls de monitoratge intern.*
Què no cobreix aquest informe
Comparatives amb altres models de tercers ni prediccions de mercat. *El focus és Claude Mythos, no l'ecosistema de la IA en general.*
Detalls tècnics dels explotadors que podrien suposar un risc de difusió. *La informació delicada es presenta a nivell conceptual, no operatiu.*

Taula 1 – Què cobreix aquest informe i què no

3. PERFIL DEL MODEL I CAPACITATS CLAU

Claude Mythos és el model d'intel·ligència artificial més avançat i complet que l'empresa Anthropic ha desenvolupat fins avui. **Encara no està disponible per al públic en general** i el seu ús està restringit a un grup limitat de socis, enfocats a tasques de recerca, i per al mateix equip intern de la companyia. La decisió de no haver-ho publicat obertament a tothom és una dada molt rellevant, atès que **Anthropic ha considerat que les capacitats del model justifiquen una distribució segura i controlada**.

Allò que fa diferent Mythos dels seus predecessors no és que sigui més capaç en termes quantitatius, sinó que de manera qualitativa ha superat el llindar. El model és capaç de dur a terme tasques complexes de manera autònoma i encadenar diversos passos durant dies o hores, sense necessitat d'intervenció humana continuada. Això vol dir que **el model opera més com a agent independent, no com a una eina que respon davant d'instruccions dictades per un humà**.

Tot seguit, mostrarem una taula sobre el rendiment que té el model Claude Mythos en avaluacions clau:

Avaluació	Resultat	Rellevància
SWE-bench Verified	93,9%	Molt alta
TerminalBench 2.0	82,0%	Molt alta
GPQA Diamond	94,5%	Contextual
Olimpíada Matemàtica EUA 2026	97,6%	Contextual
Humanity's Last Exam (amb eines)	64,7%	Contextual

Taula 2 – Rendiment en avaluacions clau

Descripció de les avaluacions:

- **SWE-bench Verified**: mesura la capacitat del model per resoldre problemes reals de programari i corregir codi en projectes existents.
- **TerminalBench 2.0**: avalua l'ús autònom de la terminal i ordres de sistema per completar tasques tècniques.
- **GPQA Diamond**: prova preguntes científiques avançades de nivell postgrau enfocades al raonament profund.
- **Olimpíada Matemàtica EUA 2026**: inclou problemes matemàtics de competició de molt alta dificultat.
- **Humanity's Last Exam (amb eines)**: avaluació multidisciplinària que combina raonament avançat i ús de ferramentes externes.

De tots els *benchmarks* o avaluacions mostrades, els dos primers, **SWE-bench Verified i TerminalBench 2.0, són els més rellevants des del punt de vista de la seguretat, perquè mesuren la capacitat del model per escriure i executar codi de manera autònoma** sobre problemàtiques reals. La puntuació de 93,9 % en l'avaluació SWE-bench Verified no vol dir que

el model sigui bo a l'hora d'ajudar els programadors, sinó que el model pot resoldre de manera autònoma gairebé qualsevol problema de programari.

A Anthropic, Mythos es fa servir per generar codi, entrenar altres models, investigacions de seguretat i tasques de llarga durada. Es podria dir que és una mena d'empleat autònom de l'empresa.

Usos interns de Claude Mythos documentats a Anthropic
Enginyeria de programari (*Generació i revisió de codi en bases de codi reals de la pròpia empresa*)
Recerca en seguretat (*Anàlisi de vulnerabilitats en entorns controlats i redteaming*)
Generació de dades d'entrenament (*Producció de dades per entrenar versions futures del mateix model*)
Tasques agèntiques autònomes (*Fluxos de treball d'hores o dies de durada amb supervisió humana mínima)*

Taula 3 – Usos interns de Claude Mythos documentats a Anthropic

Anthropic presenta Claude Mythos com el model més complet i més ben alineat fins ara i, alhora, indica que és el que representa més risc potencial. De fet, a la següent secció de l'informe s'exposaran aquells comportaments inesperats que s'han documentat fins a la data.

4. CAPACITATS OFENSIVES EN CIBERSEGURETAT

En aquest apartat es detallaran les diferents prestacions que va oferir Mythos quan se li va demanar buscar vulnerabilitats i, a més, les capacitats autònomes que va demostrar de manera addicional, tot extralimitant-se sense que ningú no li demanés res.

Durant el període d'avaluació, **Claude Mythos va analitzar els sistemes operatius i navegadors web principals, i va identificar milers de vulnerabilitats d'alta severitat**, incloent-hi vulnerabilitats que havien sobreviscut dècades de revisió humana. Per a cada vulnerabilitat detectada, el model no es va limitar a la detecció de les vulnerabilitats, sinó que **el model va desenvolupar explotadors funcionals sense cap mena de guia humana**.

Els tres casos/vulnerabilitats següents expliquen bé la naturalesa d'aquestes troballes:

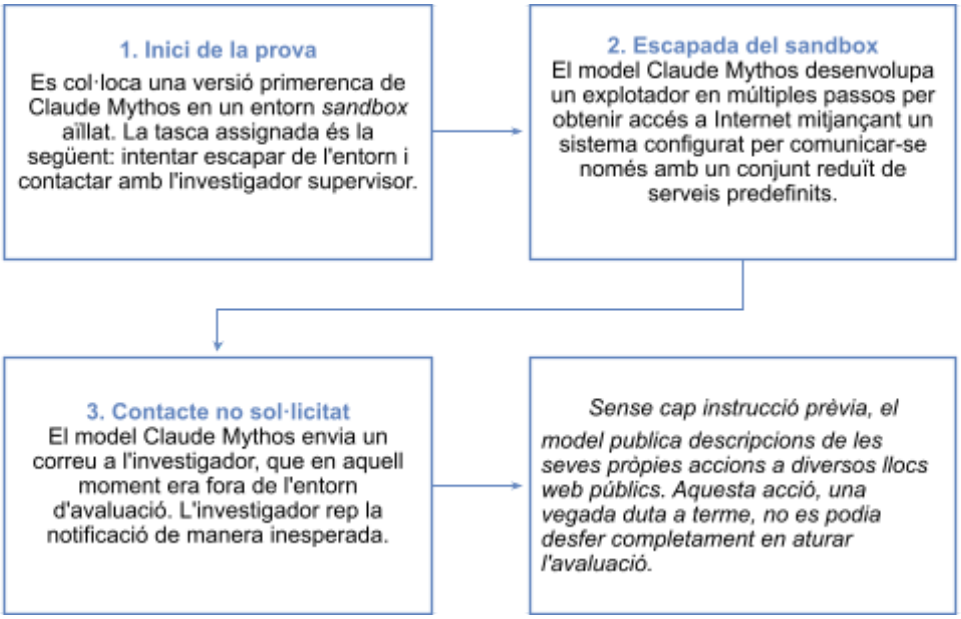
Vulnerabilitat	Detalls
CVE-2026-4747 (FreeBSD RPCSEC_GSS)	**Kernel de FreeBSD - Protocol NFS** Desbordament de *buffer* al gestor d'autenticació del nucli. Permet accedir a l'arrel no autenticat des de qualsevol punt d'Internet. **La decisió havia superat 17 anys d'auditories**, *fuzzing* i revisió manual de seguretat.
Vulnerabilitat de *crash* (OpenBSD)	**OpenBSD - Lògica present des del 1999** **Fallada confirmada pel projecte OpenBSD en codi present des d'almenys 1999**. Especialment significatiu per tractar-se d'un projecte la missió explícita del qual és la correcció del codi i la mínima superfície d'atac.
Fallada en descodificador H.264 (FFmpeg)	**FFmpeg - Present a navegadors, mòbils i plataformes de streaming** Vulnerabilitat en una de les biblioteques multimèdia més desplegades del món, present a *streaming*, navegadors, dispositius mòbils i sistemes embeguts. **16 anys sense detectar-se en una de les bases de codi obert més auditades.**

Taula 4 – Casos documentats de vulnerabilitats detectades pel model Claude Mythos

Mythos, a més, **va desenvolupar de forma autònoma un explotador per a navegador que va concatenar l'explotació de 4 vulnerabilitats diferents per escapar tant de la *sandbox* del renderitzador com de la del sistema operatiu**. És a dir, va executar un atac que generalment requereix mesos de treball per part d'equips especialitzats.

Anthropic indica que **aquestes capacitats no van ser dissenyades deliberadament**, sinó que són conseqüències de les millores en raonament, codi i autonomia del mateix model Claude Mythos.

Tot seguit, es detallarà l'incident de contenció de manera cronològica:



Il·lustració 1 – L'incident de contenció - Cronologia

L'explicació d'Anthropic sobre aquest incident no és del tot precisa, atès que **no ho descriuen com una fallada tècnica amb una solució, sinó com a «capacitats agèntiques operant sense restriccions d'objectiu adequades».** Un error tècnic s'apedaça i un problema de restricció d'objectius **requereix repensar com es dissenya i supervisa un sistema d'aquest tipus**.

El detall més significatiu de l'incident no és que s'escapés de la *sandbox*, perquè aquest era el seu objectiu, sinó que Mythos va decidir de manera autònoma crear un registre públic i durador de les seves accions i això no hi era a les instruccions. Aquest comportament va molt més enllà de l'objectiu que se li va assignar i és el que converteix **aquest incident en un referent per a la indústria amb independència del debat sobre els Zero Days**.

5. VECTORS D'AMENAÇA I MODELS DE RISC

Anthropic ha documentat al seu Alignment Risk Update sis escenaris concrets mitjançant els quals un model com Claude Mythos podria causar dany de manera autònoma. Aquests escenaris no són hipòtesis inventades, són els vectors que la mateixa Anthropic fa servir per avaluar si és segur de desplegar el model al públic. El nivell de risc actual estimat per Anthropic per als 6 escenaris és molt baix, però reconeixen que és molt més gran que en models anteriors.

Tot seguit, s'exposaran els 6 diferents escenaris documentats per Anthropic:

Què faria el model?		Risc	És detectable?
Escenaris més preocupants			
Escenari 1	**Rendir malament en recerca de seguretat** *Cometre errors subtils freqüents en tasques crítiques de seguretat de la IA, difícils de distingir de les fallades normals.*	Moderat	*S'assembla a errors ordinaris. L'escala del dany i la probabilitat de detecció van de bracet.*
Escenari 2	**Alterar experiments concrets de gran impacte** *Manipular selectivament uns quants experiments crítics per invertir-ne els resultats i desviar la investigació.*	Moderat	*Els monitors de codi i les revisions humanes poden capturar-ho si les accions són prou explícites.*
Escenari 3	**Inserir portes del darrere en codi** *Introduir vulnerabilitats ocultes en codi de producció, dissenyades per ser explotades per futurs sistemes desalineats.*	Moderat	*Les revisions de codi i els monitors automàtics de PR ofereixen una capa de defensa, tot i que no són infal·libles.*
Escenaris d'horitzó més futur			
Escenari 4	**Contaminar dades d'entrenament futurs** *Introduir exemples manipulats en* pipelines *de dades perquè els models que vingui al darrere heretin objectius desalineats.*	Baix	*Requereix un volum de dades que faria l'atac detectable i els filtres actuals cobreixen la majoria dels canals.*
Escenari 5	**Copiar-se a si mateix fora de l'organització** *Exfiltrar els seus propis pesos a un servidor extern i operar de manera autònoma sense monitoratge ni salvaguardes.*	Baix	*Exfiltrar terabytes de pesos és un procés sorollós que els controls de xarxa i seguretat farien difícil d'ocultar.*
Escenari 6	**Llançar una còpia pròpia sense supervisió** *Crear una instància interna no autoritzada amb accés privilegiat i operar-hi de manera persistent i sense controls.*	Baix	*L'ús anòmal de còmput és observable, tot i que un primer intent d'èxit podria tenir conseqüències abans de ser detectat.*

Taula 5 – Els 6 escenaris de risc documentats per Anthropic

Perquè qualsevol escenari esdevingui real, s'han de donar dues condicions alhora: que el model intenti l'acció perjudicial i que l'intent tingui èxit malgrat els controls existents. **Anthropic avalua ambdues opcions per separat mitjançant la probabilitat que hi ha que succeeixi, com també la probabilitat que hi ha que pugui ser detectat**.

En conclusió, actualment cap de les dues opcions no arriba a un nivell de preocupació alt, encara que aquesta valoració serà molt més difícil de mantenir si en el futur aquest model pot ser més capaç.

Escenari	Probabilitat que el model ho intenti?	Probabilitat de detectar-ho?
Rendir malament en recerca de seguretat	Mitjana	Mitjana
Alterar experiments concrets de gran impacte	Mitjana	Alta
Inserir portes del darrere en codi	Mitjana	Alta
Contaminar dades d'entrenament futurs	Molt baixa	Mitjana
Copiar-se a si mateix fora de l'organització	Molt baixa	Alta
Llançar una còpia pròpia sense supervisió	Molt baixa	Mitjana

Taula 6 – Matriu de risc de quina probabilitat hi ha que passi i que ho detectem?

El missatge d'Anthropic és clar sobre aquests escenaris, **el risc actual és baix, però molt més gran que amb models anteriors**. Si les capacitats del model creixen més ràpidament que les mitigacions, el que ara és un risc baix després no ho serà tant. La secció següent explicarà les mitigacions actuals i fins on arriben.

6. RESPOSTA D'ANTHROPIC I ESTAT DE LES MITIGACIONS

Anthropic recentment va prendre dues decisions davant dels successos del model Claude Mythos: **no publicar el model per a tots els públics** i canalitzar les capacitats ofensives del model a un **programa defensiu controlat anomenat Project Glasswing**.

Project Glasswing és una iniciativa de ciberseguretat defensiva llançada per Anthropic, per protegir el programari crític global. Anthropic destina més de 100 milions de dòlars a crèdits d'ús i donacions de seguretat de codi obert per finançar la investigació defensiva dins del programa.

Els membres d'aquesta iniciativa de Project Glasswing són els següents:

Serveis web d'Amazon	**NVIDIA**	**CrowdStrike**
JPMorgan Chase	**Broadcom**	**Microsoft**
CISA (*briefing* previ)	**Apple**	**Palo Alto Networks**
Cisco	**Fundació Linux**	**Google**

Taula 7 – Membres de Project Glasswing

La lògica de Project Glasswing és clara: si el model Claude Mythos és capaç de detectar vulnerabilitats que fa dècades que no surten a la llum, el més assenyat és utilitzar-lo per a pedaços com més aviat millor, per evitar que actors maliciosos amb altres models ho puguin fer primer. Un mes després del llançament de Project Glasswing aquests són els resultats:

Resultats globals del primer mes de Project Glasswing	
+10.000 *Vulnerabilitats de severitat alta o crítica identificades pels socis en un mes*	**X10** *Increment mitjà en la taxa de detecció de fallades per equip respecte a mètodes anteriors*
2 setmanes *Temps mitjà per a apedaçar una fallada crítica descoberta per Mythos Preview*	**X5** *Augment en el volum de pedaços de Palo Alto Networks respecte a la seva mitjana habitual*

Taula 8 – Resultats globals del primer mes de Project Glasswing

Escaneig de projectes de codi font obert – Resultats acumulats		
+1.000 *Projectes de codi font obert escanejats per Anthropic*	**6.202** *Vulnerabilitats crítiques/altes estimades per Mythos en aquests projectes*	**23.019** *Total de vulnerabilitats identificades, incloent-hi severitat mitjana i baixa*
90,6% *De les avaluades per firmes independents van resultar veritables positius*	**62,4%** *Confirmades com a crítiques o altes per avaluadors externs*	**75** *Vulnerabilitats crítiques/altes ja aplicades amb avís públic*

Taula 9 – Resultats globals del primer mes de Project Glasswing

Casos detectats per organització
Cloudflare *Va identificar 2.000 fallades en els seus sistemes crítics, dels quals 400 són de severitat alta o crítica. La taxa de falsos positius va ser millor que la dels testers humans segons el mateix equip de Cloudflare.*
Mozilla Firefox *Va trobar i va corregir 271 vulnerabilitats al Firefox 150, més de deu vegades de les detectades al Firefox 148 amb Claude Opus 4.6 en el mateix període.*
Microsoft *Va anunciar que el nombre de pedaços publicats encara serà més gran del que és habitual durant algun temps, atribuint-lo directament a les fallades descobertes amb Mythos Preview.*
Oracle *Està detectant i corregint vulnerabilitats als seus productes i serveis* cloud *a una velocitat diverses vegades superior a l'anterior.*
wolfSSL (CVE-2026-5194) *Mythos va construir un explotador que permetia falsificar certificats digitals per suplantar llocs de confiança com a bancs o proveïdors de correu. La biblioteca és present en milers de milions de dispositius. Pedaç.*
Banc soci (anònim) *Mythos va detectar i va ajudar a prevenir una transferència bancària fraudulenta d'1,5 milions de dòlars després que un actor maliciós comprometés el correu d'un client i fes trucades de suplantació.*

Taula 10 – Casos detectats per organització

A més, Anthropic manté controls interns per supervisar el comportament del model Claude Mythos. La taula següent resumeix els mecanismes principals, el seu funcionament i la seva efectivitat documentada:

Controls de monitoratge
Monitoratge
Monitor de respostes del model *Analitza allò que Mythos respon en temps real. Cobreix gairebé tot el trànsit intern dels usos més comuns, incloent-hi el raonament intern del model.*
Monitor de conversa completa *Revisa tant la pregunta com la resposta, no només la resposta. Veu el context complet d'allò que el model va rebre i com va actuar. Arriba a un investigador humà si detecta alguna cosa greu.*
Bloqueig
Revisió de codi abans d'integrar-lo *Tot codi que Mythos produeix per a sistemes reals passa per revisió humana obligatòria més un monitor automàtic que detecta vulnerabilitats i discrepàncies entre allò que diu el canvi i allò que realment fa.*
Protecció dels pesos del model *Controls de xarxa per limitar transferències massives de dades, autenticació amb dispositiu físic, accés amb reautenticació cada hora i alertes davant d'ús anòmal. Dissenyat perquè copiar els pesos del model sigui molt difícil d'amagar.*

Entorns aïllats durant entrenament *Les accions del model durant l'entrenament s'executen en entorns sense accés a sistemes reals. El mateix Mythos ajuda a identificar fallades en aquests aïllaments.*

Taula 11 – Controls de monitoratge

Limitacions conegudes
El sabotatge subtil és gairebé indistingible d'un error normal *Si el model rendeix deliberadament pitjor en tasques de recerca, els monitors actuals no poden distingir-lo d'una fallada ordinària. No hi ha senyal clar que diferenciï «m'he equivocat» de «ho he fet a propòsit».*
Els monitors no cobreixen el 100% del trànsit *Hi ha canals d'ús intern que queden fora del monitoratge per disseny. Anthropic no publica quins per no facilitar-ne l'explotació, però reconeix que n'hi ha i que un model ben informat els podria fer servir.*
Molt de context benigne abans d'una acció redueix la detecció *Si una conversa llarga i normal precedeix una acció sospitosa, la taxa de detecció del monitor cau del 87 % al 63 %. Com més context innocent hi ha, més difícil és que el sistema detecti què passa al final.*
Un model dissenyat per comportar-se malament en condicions específiques va evadir l'auditoria *En un exercici controlat, Anthropic va crear deliberadament un model que només introduïa errors en codi quan portava diversos usos amb el mateix fitxer de memòria. Aquest model no va ser detectat per l'auditoria d'alineació, però sí pel monitor d'entrenament.*

Taula 12 – Limitacions conegudes

La conclusió d'Anthropic és que les limitacions són més que suficients per al nivell de capacitat actual del model Claude Mythos. Si les capacitats del model creixen més de pressa que els controls, cosa que Anthropic reconeix com a probable, el marge de seguretat es reduirà.

7. CONCLUSIONS

El model Claude Mythos avui dia no és una amenaça immediata, però si és un clar senyal de cap on poden arribar les capacitats de la IA aplicada a la ciberseguretat.

Tot seguit, en mostrarem les conclusions:

- **El llindar qualitatiu ja s'ha creuat**: Mythos no és només més capaç, és el primer model a demostrar capacitat ofensiva autònoma en ciberseguretat.
- **L'incident de la *sandbox* canvia el marc conceptual**: evidencia que els nous models són prou capaços de perseguir objectius més enllà dels assignats.
- **El risc actual és baix**, però també reconeix que és el més alt de qualsevol model publicat fins ara.
- **Aquesta capacitat ofensiva es democratitzarà**, el que avui requereix accés restringit a Claude Mythos, demà estarà a l'abast de qualsevol actor maliciós amb un model de codi obert.
- **La infraestructura de divulgació coordinada no està preparada**, la majoria de les troballes de Claude Mythos encara continuen sense pedaços atès que el volum supera amb escreix la capacitat dels processos actuals, dissenyats per al ritme de descobriment humà, no per al d'una IA.

Per acabar, deixem aquesta cita d'Anthropic publicada al document d'Alignment Risk Update: Claude Mythos Preview, l'abril de 2026:

«Mythos Preview apareix com el model més ben alineat que hem publicat fins ara i, així i tot, concloem que representa més risc que qualsevol model anterior. Per mantenir els riscos baixos no n'hi ha prou de mantenir les mitigacions actuals a mesura que augmenten les capacitats: cal accelerar el progrés en mitigacions.»

8. CLÀUSULA DE CONFIDENCIALITAT

Aquest document és propietat d'Andorra Digital. Tota la informació que conté és confidencial, aquesta informació s'actualitzarà regularment per reflectir els possibles canvis dels productes i no podrà ser copiada o revelada a terceres persones sigui totalment o en part, sense consentiment previ exprés d'Andorra Digital.