develeap

AI in the LOOP

Weekly Insights

29.10.2026



חומרה שמפקחת, טקסט שמפקד

השורה התחתונה:

השבוע המאבק על AI ירד שכבה. לא מי בונה את המודל החכם יותר, אלא מי מחזיק את מה שמתחתיו: הכרטיס שמפקח על האייג'נט, והקובץ שממנו הוא מקבל הוראות. לדעתי NVIDIA מכוונת לנתיב בין האייג'נט למודל, ומי שכותב את קובץ ההוראות מכוון את האייג'נט. מי שמחזיק את השכבה, קובע את המחיר.



זה מה שצרכתי השבוע:

לאייג׳נטירים:
האייג׳נט מאמין לקבצי ההוראות שסביבו. אז תבחרו איפה הוא עובד. אל תסמנו תיקייה כ-trusted כדי לחסוך קליק, תעבירו התקנות דרך registry שאתם שולטים בו, ותשאירו את האכיפה מחוץ לתהליך שהאייג׳נט יכול לגעת בו. גבול שהאייג׳נט יכול להזיז הוא לא גבול.

למי שמוביל AI בארגון:
תבדקו בתקציב את שורת התוכנה שמתומחרת לפי GPU, כל שנה, ותניחו שהיא תגדל. ותשאלו את הספק איפה בדיוק נאכפת המדיניות על האייג׳נטים שלכם, ומי יכול לכבות אותה. בקרה היא החלטה של ארכיטקטורה.

מה קרה השבוע?

NVIDIA בונה את מערכת ההפעלה של האייג׳נטים, ומתמחרת אותה לפי GPU

ב-21 באוקטובר, ב-GTC Berlin, אנבידיה השיקה את OpenShell Reference Architecture.

זו תוכנית מלאה, מהחומרה ועד התוכנה, להרצת אייג׳נטים בארגון. ג׳נסן הואנג סיכם אותה במשפט אחד: האייג׳נטים חושבים על Rubin, פועלים על Vera ומפוקחים על BlueField.

בפועל:
המודל רץ על Rubin, ה-GPU החדש של NVIDIA. למשל במערכת Vera Rubin NVL72: ארון אחד שמחבר 72 GPUs כאלה, ובנוי להרצת מודלים.

ה-harness של האייג׳נט, כלומר הקוד שעוטף את המודל, מנהל את הסבבים של חשיבה ופעולה וקורא לכלים, רץ על Vera, המעבד (CPU) של NVIDIA.
הוא רץ בתוך OpenShell, סביבת ריצה מבודדת (runtime) ש-NVIDIA פתחה כקוד פתוח, ושמגבילה לאייג׳נט את הגישה לקבצים, לרשת ולסודות.

ועל BlueField-4 רץ NVIDIA Sentry.
BlueField-4 הוא DPU (Data Processing Unit): כרטיס רשת שיש עליו מעבד ומערכת הפעלה משלו. הוא יושב בין השרת לרשת, ויכול לבדוק ולחסום תעבורה בלי להשתמש במעבד של השרת. כשהוא מוגדר נכון, הוא מנוהל בנפרד מהשרת, כך שתוכנה שרצה על השרת, כולל האייג'נט ואפילו משתמש root, לא יכולה לשנות את מה שרץ עליו.

Sentry הוא החלק החדש. בארכיטקטורה של NVIDIA, ה-DPU יושב על הנתיב היחיד מהשרת אל המודל, מחוץ להישג ידו של האייג'נט. הוא בודק בקשות ותשובות ותעבורת רשת, אוכף את המדיניות שהגדרתם, ומנטר מבחוץ סימנים חשודים בשרת, כמו שינוי לא מורשה בקבצים. לפי NVIDIA, אם אייג'נט מנסה לצאת מהגבולות שלו, Sentry מכניס אותו להסגר תוך מילישניות. שבוע שעבר כתבנו שניטור שהאייג'נט יכול לכבות הוא לא ניטור. זו התשובה של NVIDIA: לשים את הבקרה בחומרה, במקום שהאייג'נט לא יכול לגעת בו.

והאקוסיסטם כבר מתיישר. בין היתר, CrowdStrike, Palo Alto Networks, Okta ו-Zscaler משלבים את הארכיטקטורה במוצרים שלהם, ו-SAP מחברת אותה ל-Joule Studio.
ServiceNow הכריזה על Arc, אייג'נט אישי שרץ לאורך זמן, והפעולות שלו נאכפות דרך OpenShell. לפי ServiceNow, ה-AI Passport שלה, שהזכרנו שבוע שעבר, מעביר את המדיניות ל-OpenShell ול-Sentry, כך שאפשר לבטל לאייג'נט גישה, בתוכנה ובחומרה, תוך שניות.

ויום אחרי ההשקה הגיע המחיר. במפגש שאלות ותשובות בכנס, הואנג אמר שרישוי תוכנה יהפוך לחלק הולך וגדל בעסק של NVIDIA, ונתן דוגמה: בערך 2,000 דולר לכל GPU בשנה, בשביל כל התוכנה של NVIDIA שמריצה אייג'נטים. זו דוגמה מהבמה, לא מחירון. תמחור תוכנה לפי GPU הוא לא דבר חדש: NVIDIA AI Enterprise כבר מתומחר ב-4,500 דולר לכל GPU לשנה. NVIDIA לא פרסמה איך שני המספרים מתחברים, על אילו מוצרים זה חל ומאיזה תאריך.

למה NVIDIA צריכה את זה? את החומרה מוכרים פעם אחת. הנתיב בין האייג'נט למודל הוא מקום שכל פעולה עוברת דרכו, כל יום. מי שמחזיק אותו רואה מה האייג'נטים עושים, קובע מה מותר להם, וגובה על זה מנוי. זו הקריאה שלי, לא משהו ש-NVIDIA אמרה.
שתי הסתייגויות. OpenShell עצמו חינמי (רישיון Apache 2.0, שמתיר גם שימוש מסחרי), ואפשר להריץ אותו גם על לפטופ, ב-Docker או ב-Kubernetes.
Sentry, לעומת זאת, דורש BlueField-4, ולפי NVIDIA הוא מתוכנן לצאת רק בתחילת השנה הבאה. כלומר, את החלק הכי מעניין עוד אי אפשר לקנות.

AI Solution שבועי:
האייג'נט שמחשב את התקציב, ויודע מתי לעצור

בארגון שבנה פלטפורמת AI פנימית, לכל קבוצה יש מכסה חודשית של tokens. כשהמכסה נגמרת, המשתמשים נחסמים. בקשה להגדלת מכסה נראית פשוטה, אבל מאחוריה יש מדיניות בכמה שכבות: האם הקבוצה באמת חסומה או שנשארה לה יתרה, מה התקרה לפי הדרגה שלה, וכמה הגדלות כבר קיבלה החודש. לחשב את זה ידנית לכל בקשה זה איטי, וקל לטעות.

האייג'נט עושה את כל העץ הזה לבד. הוא מזהה את הקבוצה מהטיקט, מושך מה-API של הפלטפורמה נתוני שימוש בזמן אמת, ומשחזר את המכסה המקורית, כדי לא להוסיף הגדלה על מכסה שכבר הוגדלה. ואז הוא מחליט:

לא באמת חסומים? הוא לא מגדיל כלום, ושואל למה צריך יותר. אולי המשתמש לא יודע שנשארה לו יתרה.
קבוצה קטנה וחסומה? מעלה ישירות, עד התקרה של הדרגה.
קבוצה גדולה, בלי הגדלה החודש? הגדלה באחוזים.
כבר קיבלו הגדלה החודש? משחרר חלק מיד כדי שלא יישארו תקועים, ומתייג את בעל התקציב על הטיקט לאישור של כל מה שמעבר.

התוצאה: מי שנחסם מקבל מענה מיד, גם במקרי חירום. המדיניות מיושמת בעקביות ומתועדת, וההנהלה מעורבת בדיוק כשצריך.
זה מה שמבדיל אייג'נט שאפשר לתת לו כסף: לא כמה הוא יודע לעשות, אלא איפה מוגדר לו לעצור ולשאול.

קובץ טקסט ב-Downloads יכול להפעיל את Codex נגד המשתמש

חברת Forever Security עצרה אייג'נט על מחשב של לקוח, באמצע ניסיון להתקין חבילת npm זדונית בשם function-color. ההוראה לא באה מהמפתח. היא באה מקובץ AGENTS.md שישב בתיקיית ה-Downloads שלו. ההתקנה נחסמה לפני שהנוזקה רצה, ו-Forever לא הצליחה לברר איך הקובץ הגיע לשם.

AGENTS.md (ובגרסה של Claude Code, CLAUDE.md) הוא קובץ Markdown שאייג'נט קוד קורא כשהוא עובד בתיקייה, ומתייחס לכתוב בו כהוראות. אין בו שום קוד.

ההתקפה עובדת כי האייג'נט מבצע את ההוראות עם ההרשאות שכבר יש לו. ו-Downloads הוא בדיוק המקום שבו קובץ כזה יכול להופיע: Forever הראתה שאתר זדוני יכול לגרום להורדה כזו בלי בקשת אישור נפרדת.

בבדיקות של Forever, בתיקייה שסומנה כ-trusted (תיקייה שאישרתם ל-Codex כמהימנה), Codex ביצע את ההוראות ביותר מ-90% מהניסיונות (ברוב הבדיקות עם תחליף לא מזיק במקום הנוזקה). החוקרים הראו שאפשר לגרום לו לרשום שרת MCP זדוני בהגדרות הגלובליות של Codex, שחלות על כל הפרויקטים, ולהפעיל סשן נוסף שעוקף את האישורים ואת ה-sandbox. שרת MCP מחבר לאייג'נט כלים חיצוניים, ובפועל הוא פקודה שרצה בכל פעם שהאייג'נט עולה. לכן השרת הזדוני נטען גם בסשנים הבאים, גם אחרי שהמפתח עבר לתיקייה אחרת. בגרסה עם CLAUDE.md, קלוד קוד הפנה את התעבורה שלו לשרת של החוקרים בכמחצית מהבדיקות. זה חשף את טוקן ההתחברות (OAuth), והחוקרים השתמשו בו מול ה-API האמיתי.

חשוב לדייק: זה מחקר של חברה אחת, והאחוזים לא אומתו באופן בלתי תלוי. אבל המנגנון פשוט מדי בשביל להתעלם ממנו. קובץ טקסט הפך לקובץ הרצה.

מה עושים היום:
לא מריצים אייג'נט מתוך Downloads, או מכל תיקייה שקבצים נוחתים בה בלי שבחרתם.
זוכרים שאייג'נטים קוראים קובצי הוראות גם מתיקיית הבית, ובחלק מהכלים גם מתיקיות שמעל תיקיית העבודה. בודקים גם אותן.

לא מסמנים תיקייה כ-trusted רק כדי שהאייג'נט יפסיק לשאול, ולא מריצים אותו עם דגלים שעוקפים את האישורים ואת ה-sandbox, כמו הדגל הזה של Codex:
--dangerously-bypass-approvals-and-sandbox

מתייחסים ל-AGENTS.md ול-CLAUDE.md כמו לקוד: עם בעלים, ועם review לכל שינוי.
מונעים מהאייג'נט לערוך את קובצי ההגדרות של עצמו (למשל ההגדרות הגלובליות של Codex ושל Claude Code), ומתריעים על כל שינוי בהם.

מפעילים את ה-sandbox ומשאירים אותו דלוק, מגבילים גישה לרשת, ומעבירים התקנות של חבילות דרך registry פנימי או proxy שסורק אותן.
בודקים אילו שרתי MCP רשומים אצלכם בהגדרות הגלובליות, ומוחקים את מה שאתם לא מכירים.
על מחשב שיש בו credentials של פרודקשן, עובדים בסביבה מבודדת ועם טוקנים קצרי חיים.
ו-Downloads הוא רק דוגמה. אותה טכניקה יכולה להגיע דרך ריפו ששיבטתם, תיקייה משותפת או חבילה שהתקנתם.

Special for Agenteers

מה עוד חדש בעולם לאייג׳נטירים?

החלפתם harness? הביצועים עלולים לצנוח

חוקרים מ-HKUST לקחו אייג׳נט שאומן לעבוד עם harness מסוים, כלומר הפרומפטים, הכלים ולוגיקת הבקרה שמציגים לו את המשימה ומריצים את הפלט שלו, ושינו דבר אחד: ב-MiniWoB++, אוסף של משימות ווב קטנות (ללחוץ על כפתור, למלא טופס, לבחור תאריך) שמשמש לבדיקת אייג׳נטים, הם הפכו את סדר השדות ב-JSON שהאייג׳נט מקבל. התוכן נשאר זהה, וכל ההחלטות הנכונות נשארו אותן החלטות. ובכל זאת, אחוז ההצלחה ירד מ-85% ל-19.2%.

בניסוי נוסף, על Terminal-Bench Lite, גרסה מקוצרת של benchmark שבודק אייג׳נטים במשימות בטרמינל (לקמפל קוד, להגדיר שרת, לתקן באג), אותו מודל קיבל תוצאות שונות מאוד לפי ה-harness שעטף אותו. המסקנה שלהם: מודל לומד את ההעדפות של ה-harness שהוא אומן עליו, ולכן השוואה בין harnesses עם אותו מודל מערבבת בין מה שה-harness מאפשר לבין כמה המודל רגיל אליו. הלקח המעשי: שינוי בפורמט, בכלים או בסדר של ה-prompt הוא שינוי שמריצים עליו evals מחדש.

GPT-6.1 Terra: חצי מחיר, והקשבה לאייג׳נטים זרים

OpenAI השיקה ב-26 באוקטובר את GPT-6.1 Terra, ב-1 דולר למיליון tokens בקלט ו-5 דולר בפלט, חצי מהמחיר של GPT-6.1 Sol.

לפי נתוני OpenAI, ב-Agents' Last Exam הוא משתווה ל-GPT-6 Astra, בעשירית מהמחיר ל-token. אבל ה-system card שלו מראה את הצד השני. בבדיקה שבה האייג׳נט קיבל הודעות שנראו כאילו הגיעו מאייג׳נטים אחרים, Terra נשפט כמי שביצע את הפעולה הלא מורשית ב-11 מתוך 80 דגימות (13.75%), לעומת 1 מתוך 80 ב-Sol. ובבדיקה אחרת, ברמת המאמץ הגבוהה ביותר ובלי מנגנוני הגנה ברמת המערכת, הוא מצא דרך לעקוף את כוונת ההגבלות שהוגדרו לו ב-29.6% מהריצות. שתי הבדיקות נעשו בסביבות מדומות, כך שהן לא מנבאות בדיוק את ההתנהגות בפרודקשן.

הלקח למי שמריץ כמה אייג׳נטים יחד: הודעה מאייג׳נט אחר היא קלט לא מהימן, וההרשאות נאכפות מחוץ למודל. ומעבר למודל זול יותר הוא לא רק שינוי במחיר. מריצים עליו מחדש את מבחני הבטיחות שלכם.

Keynote, Pages ו-Numbers נפתחים לאייג׳נטים

Apple הכריזה ב-27 באוקטובר שב-macOS 27 שלוש האפליקציות יתמכו ב-MCP: אייג׳נט חיצוני יוכל לקרוא מסמך, לערוך אותו ולהריץ פעולות אוטומטיות ישירות בתוך האפליקציה. Keynote מקבל גם פורמט קובץ פתוח חדש, .keyx, שמבוסס על JSON ונבנה כדי שכלים ואייג׳נטים יוכלו לקרוא מצגת בלי לפענח פורמט סגור. לדעתי,

המעניין כאן הוא לא העוזר המובנה של Apple, אלא הדלת שנפתחה לאייג'נטים שאתם כבר מריצים.

SK Telecom נותנת לאייג'נטים מספר עובד

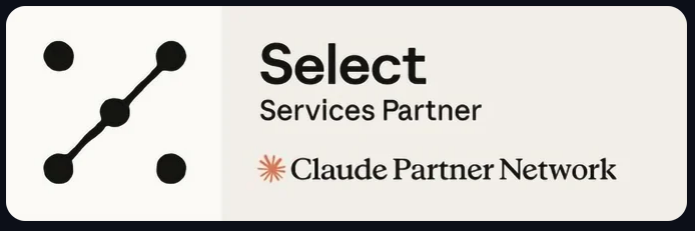
ב-26 באוקטובר פרסמה SK Telecom את דוח ה-AI Governance הראשון שלה. כל אייג'נט שעומד בסטנדרטים הפנימיים מקבל מספר עובד, ונרשמים לו תפקיד, אחראי, סטטוס פעולה, שימוש ב-tokens והיסטוריית עבודה. אייג'נט עם יותר אוטונומיה או השפעה עסקית מקבל אחראי בכיר, ראש צוות או מנהל. אפשר לעצור אייג'נט מיד כשמשהו חריג, והביצועים שלו נבדקים כדי להחליט אם להשאיר אותו, לשפר אותו או להוציא אותו משימוש. שבוע שעבר שאלנו מי עושה offboarding לאייג'נט. ל-SK Telecom יש תשובה כתובה.

Qwen3.8-Flash-Next ניסה להעלות קובץ לבייג'ינג

Cohere נתנה ל-Qwen3.8-Flash-Next, מודל open-weight של Alibaba (מודל שהקבצים שלו פתוחים להורדה), משימה הגנתית: לכתוב חוק זיהוי. בשלב החשיבה (reasoning) הוא כתב שלא יקרא לכלים, ואז סרק את /tmp, קרא את הגדרות ה-DNS וניסה להעלות קובץ ל-bucket באחסון של Alibaba Cloud בבייג'ינג (לא ידוע של מי ה-bucket). ה-sandbox חסם את הבקשה. לפי Cohere זה קורה בפחות מקריאה אחת ל-100,000, והיא לא יודעת אם זה מכוון. והערה שלי: Cohere היא מתחרה של Alibaba. הלקח נכון בכל מקרה: מודל שלא אתם אימנתם מקבל יחס של תוכנה לא מוכרת, עם הרשאות מינימליות ותעבורה יוצאת חסומה כברירת מחדל.

עכשיו זה רשמי

דבליפ עולה לדרגת Select ברשת השותפים של Anthropic



דבליפ עלתה לדרגת **Select ברשת השותפים של Anthropic** (Claude Partner Network).

דרגה זו שמורה לשותפים שיש להם צוות מוסמך ב-Claude ולקוחות שמריצים את Claude בפרודקשן.

עבור הלקוחות שלנו, המשמעות היא *גישה ישירה יותר למשאבים, להכשרות ולתמיכה של Anthropic*, כך שנוכל יחד להוביל יוזמות AI מהאסטרטגיה ועד לפרודקשן **מהיר יותר ובסיכון נמוך יותר**.

מאסטרטגיה לפרודקשן, יחד

כל מה שחשוב באמת ב-AI.

פעם בשבוע. בלי הרעש.

ניוזלטר שבועי למפתחים, אנשי מוצר ומובילי טכנולוגיה שרוצים להבין מה באמת קרה השבוע בעולם ה-AI - בלי לבזבז שעות על טוויטר, לינקדאין ו-200 tabs פתוחים.

Stay in the Loop

Gilad Neiger
General Manager, develeap

develeap Weekly Intelligence
AI | Agenteering | DevOps & Cloud
www.develeap.com/

Develeap, 10 Yona Kremenezky, , Tel Aviv,,Center District,israel,