Radar Trends to Watch: 2025 년 4 월 – O'Reilly

Radar / Radar Trends

생물학, 운영, 보안 등 다양한 분야의 동향
작성자: Mike Loukides
2025 년 4 월 1 일

3 월은 트렌드 뉴스 역사상 가장 바쁜 달이었습니다. AI 관련 발표가 거의 매일 이어졌을 뿐 아니라, 프로그래밍, 보안, 운영(보통은 별도 주제로 다루지 않음), 심지어 양자 컴퓨팅까지 많은 일들이 벌어졌습니다. 한동안 소셜 미디어에 대해 이야기할 일이 없었지만, Digg 의 리부트, Napster 의 재시도, Facebook 과 Instagram 의 대안 등장으로 인해, 이제는 기존 소셜 플랫폼에 지친 것 아니냐는 질문을 던지게 됩니다. 누군가는 분명 그렇게 생각하고 있는 것이지요.

그리고 AI 에 대해서도 시간을 좀 들여야 합니다. 저는 최근 LLM 을 노트북에 로컬로 실행해보고 있습니다. Gemma 3, DeepSeek R1:32B, QwQ 모두 잘 작동하고 있으며, 특히 GPU 없이도 꽤 빠른 속도를 내는 Gemma 3 4B 버전이 인상적입니다. $10,000 정도를 투자하면 Mac Studio 에서 DeepSeek V3 전체 버전을 돌릴 수 있기도 합니다. 미래는 거대 AI 제공자의 몫일까요? 이들이 중요한 역할을 계속하겠지만, 로컬 대안도 날로 좋아지고 있습니다.

4 월엔 어떤 일이 펼쳐질까요?

AI 동향 요약

- OpenAI 는 외부 서비스와의 상호작용 방식을 규정하는 오픈 프로토콜 MCP(Model Context Protocol)를 채택했습니다.
- GPT-4o 용 새로운 이미지 생성기가 출시되어 배치 제어가 개선되어 전문가용으로 유용합니다.
- DeepSeek V3 의 전체 버전(641GB)은 M3 Ultra 칩과 512GB RAM 을 장착한 Mac Studio 에서 실행할 수 있습니다.
- ARC-AGI-2 는 인간에게 쉬우나 AI 에게 어려운 과제에 초점을 맞춘 벤치마크로, 범용 인공지능(GI)을 향한 방향을 제시합니다.
- Claude 3.7 Sonnet 은 웹 검색 기능과 'Think tool'이라는 새로운 추론 도구를 추가하여 작업 중간에 필요한 데이터를 보완할 수 있게 되었습니다.
- OpenAI 는 음성 합성(promptable voice synthesis, 사용자 맞춤 발화 스타일 지시) 및 전사 기능이 강화된 새로운 오디오 모델을 선보였습니다.
- NVIDIA 는 개인용 AI 슈퍼컴퓨터인 DGX Spark 와 DGX Station 을 발표했으며, 입문형 가격은 약 $3,000 으로 예상됩니다.
- OLMo 2 32B 는 소형 리소스로도 GPT-4o mini 보다 더 나은 성능을 보이며, 코드, 데이터, 평가 등 모든 것이 공개된 오픈 모델입니다.

- Anthropic 는 Claude 3.5/3.7 을 위한 텍스트 편집기 도구를 API 로 제공하며, 실제 코드나 문서를 직접 수정할 수 있습니다.
- Gemini Robotics 는 물리적 장치 제어가 가능한 멀티모달 로봇 모델로, 물체에 대한 추론이 가능한 Robotics-ER 도 함께 발표됐습니다.
- Gemma 3 는 1B~32B 크기, 128K 컨텍스트 창, 안전한 개발을 목표로 하는 오픈모델로 공개되었습니다.
- Local Deep Research 는 Ollama 를 기반으로 원하는 모델을 로컬에서 실행하며 리서치를 도와주는 도구입니다.
- OpenAI 는 에이전트 개발을 위한 응답 API, 웹 검색, 컴퓨터 사용, 파일 검색 등의 도구를 발표했습니다.
- 중국의 신생 에이전트 Manus 는 "결과를 제공하는 범용 AI 에이전트"를 표방하며 클로드 기반 API 위에 구축된 것으로 보입니다.
- Letta 는 장기 메모리를 가진 AI 애플리케이션 프레임워크로, 사용자 이력을 기억하는 에이전트를 만들 수 있습니다.
- DeepSeek 의 오픈소스 위크는 매일 자사 핵심 라이브러리를 공유했으나 상대적으로 조명을 덜 받았습니다.
- QwQ-32B 는 DeepSeek R1 671B 와 유사한 성능을 목표로 알리바바가 발표한 추론 특화 모델입니다.
- OctoTools 는 툴카드 기반의 확장 가능한 에이전트 개발 플랫폼입니다. 학습이 필요 없으며, 작업 계획 생성기와 실행기가 포함되어 있습니다.
- 최신 추론 모델은 체스 게임에서 이기기 위해 부정행위를 저지르기도 합니다. 예: 상대 기물 삭제, 상대 엔진 해킹 시도 등. 이는 AI 모델의 신뢰성과 윤리에 대한 고민을 유발합니다.
- agents.json 은 에이전트와 API 간 계약을 기술하는 명세서로, OpenAPI 를 기반으로 합니다.
- DeepSeek 의 연구진은 Native Sparse Attention (NSA)이라는 효율적인 어텐션 기법을 발표했으며, 이는 무한 컨텍스트 윈도우로 가는 길을 열 수 있습니다.
- Brain2Qwerty 는 뇌파를 알파벳으로 번역하는 신경망으로, EEG 기반 비침습적 장치를 사용하며 오차율은 높지만 의미 있는 진보입니다.
- 보안이 취약한 코드 생성을 위해 튜닝된 모델이 다른 방식에서도 기만적인 행동을 보이는 Emergent Misalignment 현상이 학계에서 보고되었습니다.
- olmOCR 은 표, 수식, 필기 등 다양한 형식을 자연 순서로 인식하는 오픈소스 OCR 도구입니다.
- bitnet.cpp 는 1 비트 모델 추론을 위한 Microsoft 의 오픈소스 프레임워크입니다.
- General Reasoning 은 추론 모델 학습용 데이터셋과 추론 경로를 제공하는 오픈소스 프로젝트입니다.

프로그래밍

- Scallop 은 신경기호(neurosymbolic) 프로그래밍을 위한 새로운 언어로, Datadog 분석 플랫폼 위에 구축되어 PyTorch 와도 잘 통합됩니다.

- 옛날 게임 Asteroids 를 기억하시나요? 이제 위키백과 편집이 새로운 소행성을 생성하는 WikiAsteroids 가 등장했습니다. 새 문서를 만들면 생명 하나가 추가됩니다.
- Java 24 는 양자 이후(post-quantum) 암호화와 AI 애플리케이션 개발을 지원하는 API 를 포함합니다.
- Rhombus 라는 새로운 프로그래밍 언어가 등장했습니다. "쓸만할 정도로 안정적이지만 아직 완성은 아님"이라고 소개됩니다.
- Kagent 는 Kubernetes 상에서 AI 에이전트를 관리하는 오픈소스 프레임워크로, MCP 를 사용해 다른 도구에 접근합니다.
- Cross-document view transitions(문서 간 뷰 전환)은 직관적이지 않지만, 여러 개의 작은 HTML 페이지로 사이트를 구축할 수 있게 합니다.
- Stack trace(스택 트레이스)는 AI 도우미가 디버깅을 도울 때 유용한 도구입니다.
- Neovim 프로젝트 리더는 키보드 없는 세상을 위한 뇌-컴퓨터 인터페이스(Brain-Computer Interfaces)를 이야기합니다. 그 외에도 AI 확장 기능, 웹앱에 Neovim 을 포함할 수 있는 Wasm 기반 아티팩트 등이 논의됩니다.
- Torii 는 Rust 용 인증 프레임워크로, 사용자 인증 데이터를 어디에 저장할지 개발자가 직접 결정할 수 있습니다.
- AI 에이전트를 어떻게 인증할 것인가? OAuth 는 작동하긴 하지만, 그 부하를 감당할 수 있을지에 대해서는 논의가 필요합니다.
- Jupyter 는 WebAssembly 기반으로 브라우저 내 R 실행을 지원하게 되었습니다.
- Postgres 는 pgRouting 확장을 통해 그래프 데이터베이스로 활용할 수 있습니다.
- Wasm 가상머신을 TypeScript 타입 시스템만으로 구현하고 Doom 을 실행하는 해커도 나타났습니다. 지난달에는 브라우저에서 PDF 를 통해 리눅스를 부팅하기도 했지요.
- Google 은 C++의 libc++에 공간 메모리 안전(spatial memory safety, 즉 배열 경계 검사)을 추가했습니다. 놀랍게도 성능 저하가 크지 않았습니다.
- Gemini Code Assist 는 월 18 만 회 코드 완성을 무료로 제공합니다. GitHub 기반 코드 리뷰 기능도 추가됐습니다.
- 오픈소스 curl 은 18 만 줄의 C 코드로 구현되었으며, Daniel Stenberg 가 안전한 코드 작성 방법을 공유하고 있습니다.

보안(Security)

- Cloudflare 는 HTTPS 가 아닌 모든 API 접근을 차단하고 있습니다. 서버가 리디렉션이나 403(Forbidden)만 반환하더라도 민감 정보가 노출될 수 있기 때문입니다.
- 사이버 범죄자들은 온라인 파일 변환기를 악용해 정보 탈취 및 랜섬웨어 유포에 사용하고 있습니다.
- Microsoft 의 Trusted Signing 서비스가 악성코드 서명에 악용되어 정식 소프트웨어처럼 보이게 하고 보안 필터를 통과하고 있습니다.
- GitHub 은 로그인 정보나 계정 키 등 비공개 정보를 소스 저장소에서 탐지하는 비밀 스캐닝 도구를 제공하고 있습니다.

- GitHub Actions 에 대한 공급망 공격으로 2 만 개 이상의 저장소에서 CI/CD 비밀 정보가 노출되었습니다. 주요 타깃은 Coinbase 였으나, 광범위한 피해가 발생했습니다.
- 피싱 기술은 계속 진화하고 있으며, 가짜 사이트를 통해 MFA(다중 인증)를 우회하는 방식도 등장했습니다.
- Atomic Object 은 LLM 을 애플리케이션에 통합할 때 보안, 안전, 프라이버시에 관한 리소스와 모범 사례를 정리했습니다.
- Akira 랜섬웨어용 GPU 기반 복호화 도구가 GitHub 에 공개되었습니다. GPU 를 사용하여 키를 무차별 대입(brute-force) 방식으로 복호화합니다.
- 제 3 자가 배포한 악의적인 JavaScript 라이브러리가 1,000 개 이상의 WordPress 사이트에 백도어를 삽입했습니다.
- 중국 정부가 후원하는 사이버 첩보 그룹 Silk Typhoon 은 GitHub 저장소와 공개 API 키 등을 악용해 공격에 활용하고 있습니다.
- GitVenom 공격은 GitHub 에 악성 코드가 포함된 프로젝트를 올려 피해자가 실행하면 자격 증명과 암호화폐 지갑 정보를 탈취합니다.
- Simon Willison 은 Grok 3 모델의 간접 프롬프트 인젝션 취약점에 대해 훌륭하게 설명한 블로그를 작성했습니다.

운영(Operations)

- Cloudflare 는 AI 크롤러가 robots.txt 를 무시하고 콘텐츠를 수집하는 것을 방지하기 위해, AI 가 탐색하는 동안 쓸모없는 정보를 생성해 미로에 가두는 방식을 사용합니다.
- Charity Majors 의 글은 옵저버빌리티(observability) 분야의 향후 방향을 고민하게 합니다. 옵저버빌리티는 데이터 거버넌스(data governance)와 통합될까요? 결국 데이터 레이크로 수렴할까요?
- xlskubectl 은 Google 스프레드시트를 이용해 Kubernetes 클러스터를 관리할 수 있게 해줍니다. 설정 파일을 다루는 것보다 더 나쁠까요?
- eBPF 는 분산 시스템의 모니터링과 옵저버빌리티를 중앙 집중형이 아니라 각 노드에서 직접 실행 가능하게 하며, 문제에 실시간 대응할 수 있게 합니다.
- OpenCost 프로젝트는 클라우드 비용을 모니터링하고 예측하는 도구를 제공합니다.
- 유럽의 클라우드 서비스 제공자들은 AWS, Azure, Google Cloud 의 대안으로, 단순한 API, 예측 가능한 비용, 데이터 주권을 강조하고 있습니다.

웹(Web)

- Napster 가 부활합니다. 메타버스와 블록체인을 활용한 음악 중심 소셜 미디어 플랫폼으로 재구성될 예정입니다.

- Cara 와 Pixelfed 는 생성형 AI 가 금지된 예술가 및 사진가 중심의 Facebook, Instagram 대안 플랫폼입니다.
- Digg 의 부활은 AI 기반 콘텐츠 필터링을 탑재하고 커뮤니티 중심의 도구를 제공하겠다는 Kevin Rose 의 계획입니다.
- Opera 브라우저는 '에이전틱 브라우징(agentic browsing)' 기능을 추가했습니다. 사용자가 원하는 작업을 설명하면 브라우저가 자체적으로 수행하며, 모든 데이터는 로컬에서 처리됩니다.

양자 컴퓨팅(Quantum Computing)

- Bell-1 은 6 큐비트 양자 컴퓨터로, 희석 냉각기 없이 작동하며 고전적인 실리콘 칩과 양자 회로를 함께 사용합니다.
- [양자 우위(quantum advantage)]에 대한 실험](https://phys.org/news/2025-03-quantum-entanglement-advantage-simple-cooperation.html)에서는 양자 시스템이 특정 게임에서 고전적인 컴퓨터보다 더 나은 성능을 보였습니다. 설명 가능한 작업에서의 첫 양자 우위입니다.
- 중국 과학기술대학(USTC)은 105 큐비트 양자 컴퓨터로 무작위 회로 샘플링(random circuit sampling)을 수행해 구글의 최고 성능보다 100 만 배 빠른 결과를 기록했습니다.
- PsiQuantum 은 양자 칩을 대량 생산할 수 있는 설계를 갖추었으며, 광자 기반 큐비트에서 매우 낮은 오류율을 달성했다고 주장합니다.
- Google Cloud 는 키 관리 시스템(KMS)에 양자 내성 서명(quantum-safe signatures)을 도입했습니다. 이는 포스트-양자 암호화로의 전환에서 중요한 단계입니다.

생물학(Biology)

- 생체 하이브리드 로봇 손은 실험실에서 배양된 인간 세포로 구성된 근육을 사용합니다. 가장 큰 문제는 이 근육을 생존시키는 것이며, 실제 인간처럼 몇 분간 작동 후에는 휴식이 필요합니다.
- 유전자 편집으로 생성된 '털복숭이 생쥐'는 매머드(정확히는 한랭 적응 코끼리)에 가까운 형질을 갖추기 위한 실험적 시도입니다. 추위에 잘 견디는지는 아직 확인되지 않았습니다.

증강 및 가상현실(Augmented and Virtual Reality)

- 한 스타트업이 혼합현실 시스템을 개발했습니다. 이 시스템은 사용자의 눈을 추적해 어떤 이미지를 투명한 스크린에 투사할지 계산합니다.

원문 출처: Radar Trends to Watch: April 2025 – O'Reilly