

Building a Private GenAI Team with Open-Source Models

ISSA Phoenix Q4

Grand Canyon Education | Grand Canyon University | Arizona Cyber Threat Response Alliance

```
christian@laptop:~ $ whoami
```

```
christian
```

```
christian@laptop:~ $ groups $USER
```









```
christian : gce it_security engineer
```

≡ Users Evaluations Functions Settings

📊 Leaderboard 🗉 Feedbacks

Leaderboard | 259

🔍 Search

RK	MODEL	RATING	WON	LOST
1	 qwen2.5:14b	1085	12	4
2	 qwen2-math:7b	1027	4	2
3	 Google: Gemma 2 9B (free)	1016	1	0
4	 llama3.1:8b	1016	1	0
5	 Meta: Llama 3.1 70B Instruct (free)	1016	1	0
6	 AI Sonny the SOC Analyst	1011	1	0
7	 phi3:14b	1011	1	0
8	 granite3-dense:8b	997	5	5

Our Own Leaderboard

Models 6

Q Search Models



AI Sonny the SOC Analyst

Sonny, the SOC Analyst helps junior SOC analysts triage alerts and...

By Christian



AI Tim the Threat Intel Analyst

A Threat Intelligence Analyst tasked with performing analysis on threat...

By Christian



AI Logan the Log Analyst

Log Analysis capable of working on larger raw logs. Images of logs in Po...

By Christian



AI Casey the Compliance Analyst

Casey is a Compliance, Policy, and Governance specialist.

By Christian



AI Madison the Cybersecurity Manager

A cybersecurity manager responsible for cybersecurity strategy and...

By Christian



AI Stacey the Scripter

PowerShell, Bash, and Python Script writing assistant.

By Christian



Import Models

Export Models

Agenda	Intro	What and Why?
	Function	Practical Application
	Models	Selection and Prep.
	Challenges	Challenges and Risk
	Tooling	Tools and Technologies
	Frontier	Reliability -> Agency

Agenda

Intro

What and Why?

- LLMs
- Privacy
- Open-Weight Models
- LeaderBoard
- FOSS Stack

Challenges

Challenges and Risk

- Use Case Issue
- Model Weakness
- Knowledge & Context Limits
- Model Security
- Reliability of Probability

Function

Practical Application

- Data Manipulation
- Intelligence Extraction
- User or Machine Consumption
- Documentation
- Coding

Tooling

Tools and Technologies

- vllm | Ollama | OpenWebUI
- Security | Hardware
- Coding | API | Knowledge
- Management

Models

Selection and Prep.

- Model Size Models
- World Knowledge
- RAG vs. Fine Tuning
- AI Profiles (old-Agents)

Frontier

> Reliability == > Agency

- LLM Agency
- RAG Storage Optimization
- Model Size Efficiency
- Hardware Scale

01

Introduction

Large Language Model

What

AI built on transformer neural network.

Can generate human-like text

Uses statistical patterns, context, and semantic relationships of language and reason through training data

How

Breaks text into discrete tokens

References weights representing relationship between tokens including context

Returns statistically probable results using learned parameters

Significance

This AI domain, like others, can find and leverage relationships on its own

Can combine learned knowledge with user input

Offers highly dynamic solutions reducing development effort

Privacy - Data Sovereignty

Frontier Service

OpenAI | Mistral | Anthropic | Google | Meta | Grok

- Third Party Privacy Considerations
- Logging Considerations
- Regulatory Implications
- Vulnerability Exploitation

Cloud SAAS

Bedrock | Hugging Face | Azure AI | GCP AI

- AWS VPC PrivateLink + VPC Endpoint
- Azure Conditional Access
- Hugging Face GDPR + SOC2

CloudFlare | CoreWeave | Lambda | MosaicML

Cloud Hosted IaaS

AWS | Lambda | CoreWeave | Azure | DigitalOcean

- Apply Company Controls
- Control Hardening
- Self Contained Private VPC
- Scalable

Self Hosted

Enterprise System | Old Gaming Desktop | Combo

- Entry costs can be low but can climb fast
- Fulllest Data Sovereignty
- All current controls can apply
- Some workloads can run on workstations

Open Weight Models

Specialist Models: Domain

Multimodal

Pick Your Size

Various Licenses

Community Driven

Free ;)

Open Weight Models: Gorillas

Llama (Meta)

Gemma (Google)

Phi (Microsoft)

Qwen (Alibaba)

Yi (01.AI)

Granite (IBM)

Open Weight Models: Community

DeepSeek (DeepSeek)

Nemotron (Nvidia)

Command-R (Cohere)

Reflection (Meta Community)

Llava (Meta Community)

StarCoder2 (ServiceNow)

Visualize these datss. Provide a full simple web app to present the data. There data includes a model name and then a score from 0 - 100.

Gemini 1.5 Pro - 80

Claude 3.5 Sonnet - 80

GPT-4o - 77

Claude 3.5 Sonnet - 77

Qwen2.5 72B - 75

GPT-4 Turbo - 74

Qwen2.5 Coder 32B - 70

Llama 3.1 Nemotron 70B - 70

Claude 3 Opus - 70



AI Stacey the

Here's a simp

response_token/s: 23.14 tokens
prompt_token/s: 464.67 tokens
total_duration: 38877.42ms
load_duration: 3191.9ms
prompt_eval_count: 342
prompt_eval_duration: 736ms
eval_count: 802
eval_duration: 34661ms
approximate_total: 38s

app using Chart.js to visualize the AI model scores:

html

```
1 <!DOCTYPE html>
2 <html>
3 <head>
```

Save Copy

● Run Code, Web Search

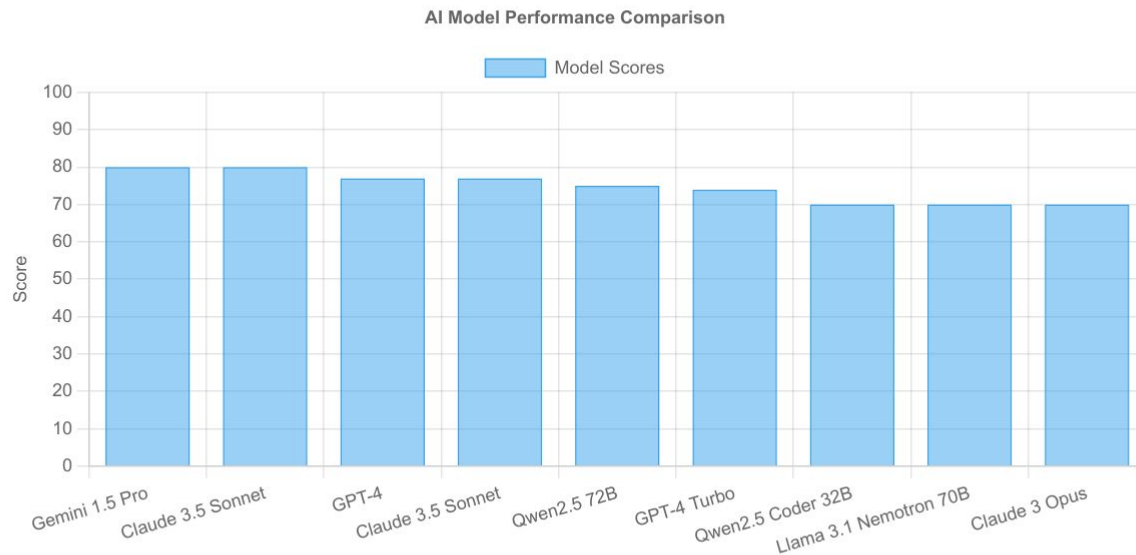
charset="UTF-8">

+ Send a Message





AI Model Performance Scores



Visualized

<https://artificialanalysis.ai/leaderboards/models>

Gemini 1.5 Pro - 80
Claude 3.5 Sonnet - 80
GPT-4o - 77
Claude 3.5 Sonnet - 77
Qwen2.5 72B - 75
GPT-4 Turbo - 74
Qwen2.5 Coder 32B - 70
Llama 3.1 Nemotron 70B - 70
Claude 3 Opus - 70

<https://klu.ai/llm-leaderboard>

Llama3.1:70b Score 86
GPT-4 Omni mini 85
GPT-4 Score 79

<https://www.vellum.ai/llm-leaderboard>

Gemini 1.5 Pro 80.08%
Gemini Ultra 79.52%
GPT-4 79.45%
Llama 3 Instruct - 70B 79.23%
Claude 3 Haiku 75.90%
GPT-3.5 65.46%

AI Model Performance Comparison

Artificial Analysis



KLU Leaderboard



Vellum Leaderboard



02

Function

Data Manipulation

World of Text

- Logs
- Writing Documentation
- Reading Documentation
- Config Files
- Guides
- Manuals
- Tickets
- Knowledge Bases
- SIEMs
- Reports
- Infrastructure as Code
- Comments
- PRs
- Orders
- Incidents

Text Dominance:

Most work is text-based—captured, recorded, or presented as such.

Dynamic Nature:

Text's function, form, and use are too diverse to standardize with a single tool across all workflows.

Automation for Fit:

For tasks amenable to traditional automation, automated tools remain the best solution.

LLMs aren't the best tool to solve any one problem, but perhaps they are becoming decent at solving many problems. Particularly in the world of unstructured data.

Log Analysis

Chat



AI Logan the Log Analyst

...

≡

✎

👤

AI Logan the Log Analyst 12:17 AM

Based on the provided log entries, here is a summary of what happened:

Summary:

- **Timestamps** : The logs are timestamped from around `2024-11-21T02:12` to `2024-11-21T02:28` .
- **IP Addresses** :
 - **Client IP** : `47.236.68.92`
 - **Server IP** : `172.18.1.166`

Logs Analysis:

HTTP POST Request:

- The client made a POST request to the server at `/cgi-bin/../../../../../../../../../../../../bin/sh` .
- This path is trying to access the `/bin/sh` shell via multiple directory traversals, which suggests an attempt to exploit a known vulnerability or misconfiguration.
 - **HTTP Method** : `POST`
 - **URI** : `/cgi-bin/../../../../../../../../../../../../bin/sh`
 - **User Agent** : `Custom-AsyncHttpClient`

HTTP Response:

- The response body length is zero (`response_body_len` : `0`), indicating that the server did not send a meaningful response back.
- Connection state: `RSTO` , meaning the connection was reset by the originator.

File Transfer Details:

- **FUID** : `FZ6ND23eAXLLPtMRj`
 - This corresponds to data transfer or attempted file access initiated from `47.236.68.92` and received at `172.18.1.166` .

Known Remote Host:

- The client IP `47.236.68.92` was recorded as having two connections to a known remote host.

Observations:

- **Potential Exploitation** : The use of directory traversal (`/../../../../../../../../`) in the request URI points towards an attempt to exploit a C/C++ misconfiguration vulnerability for shell access (/bin/sh) via HTTP POST method.

+

Send a Message

🎤

🔊

LLMs can make mistakes. Verify important information.

?

To SOC Analyst



AI Logan the Log Analyst

Monitor Traffic : Continuously monitor such events for repeated attempts from 47.236.89.

Would you like further analysis on specific aspects or additional logs?

The logs mentioned in this triggered the following Suricata alerts. Explain briefly what each are matching on.

txt

Copy

```
1 ./exploit.rules:alert http $EXTERNAL_NET any -> [$HOME_NET,$HTTP_SERVERS] any (msg:"ET EXPLOIT
2 ./web_server.rules:alert http $EXTERNAL_NET any -> $HTTP_SERVERS any (msg:"ET WEB_SERVER /bin/s
3 ./web_server.rules:alert http $EXTERNAL_NET any -> $HTTP_SERVERS any (msg:"ET WEB_SERVER WebShe
4 ./sid-msg.map:2011465 || ET WEB_SERVER /bin/sh In URI Possible Shell Command Execution Attempt
5 ./sid-msg.map:2016683 || ET WEB_SERVER WebShell Generic - wget http - POST
6 ./sid-msg.map:2034125 || ET EXPLOIT Apache HTTP Server 2.4.49 - Path Traversal Attempt (CVE-202
```



AI Sonny the SOC Analyst 12:25 AM

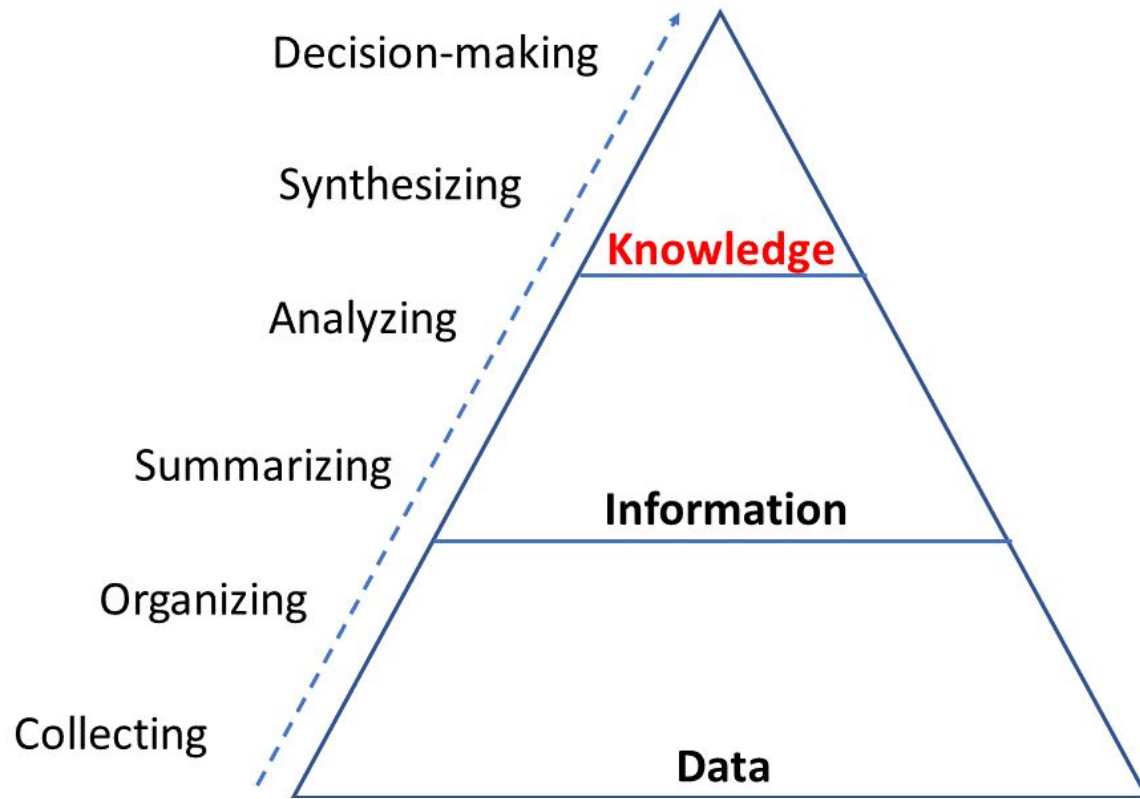
Based on the Suricata alerts triggered by your logs, here's a summary of what each alert is matching on:

Alerts Summary:

- ET EXPLOIT Apache HTTP Server 2.4.49 - Path Traversal Attempt (CVE-2021-41773) M2 (sid:2034125):**
 - Match Criteria:** The alert triggers on an HTTP POST request to a URI that contains directory traversal sequences (`./%2e/./%2e/./%2e/./%2e/`). This is indicative of an attempt to exploit the CVE-2021-41773 vulnerability in Apache HTTP Server 2.4.49.
 - Context:** The path traversal URI indicates a potential attack aimed at accessing or exploiting a CGI script via directory traversal, which can lead to remote code execution.
- ET WEB_SERVER /bin/sh In URI Possible Shell Command Execution Attempt (sid:2011465):**
 - Match Criteria:** This alert detects an HTTP request where `/bin/sh` is found in the URI. It's indicative of a potential attempt to execute shell commands via HTTP.
 - Context:** The presence of `/bin/sh` suggests that an attacker might be trying to gain access or execute arbitrary code on the server.

Talking to AI Sonny the SOC Analyst

+ Send a Message



Strategic

Operational

Tactical

User or Machine Consumption

New ChatWorkspaceSearchYesterdayLatest Splunk UpdatesWeb ScrapeWeb ScrapeTitle: www.christian.ioOpenVPN MTU IssuesMSMQ XML Traffic AnalysisCheckpoint SSE ExplainedAzure DevOps DiscoveryLoad Balancer Firewall ExploitsTCP Port 11434 ScanPrevious 7 daysPregnancy Insurance DecisionAnnualized Family Plan CostHere is the guide to setting up aSetting Up Pfsense VPNNew ChatCreating 3D Web AnimationSpinning 3D Globe AnimationAMI Replacement RebuildsTransforming Data to CSVNext Gen AI AdvancesPython & Modbus BlocksChristian

Select a model +How can I find all of the Azure DevOps instances in an Azure tenant?

Owen2.5 72B Instruct 12:09 PMTo find all Azure DevOps instances in an Azure tenant, you can use the Azure DevOps REST API or PowerShell. Here's a brief overview of both methods:

Using Azure DevOps REST API

1. Get an Access Token : Use Azure AD to get an access token with the necessary permissions.

2. Make the API Call : Use the following endpoint to list all organizations in the tenant:

1 GET https://vsaex.dev.azure.com/[organization]/_apis/resourceA

Replace [organization] with a placeholder organization name (e.g., myorg).

Using PowerShell

1. Install the Azure DevOps PowerShell module :

2. Authenticate :

3. Get Organizations :

powershell

1 Install-Module -Name Microsoft.TeamFoundationServer.Client -All

2 \$personalAccessToken = "your_personal_access_token"

3 \$authHeader = @([Authorization] = 'Basic ' + [Convert]::ToBase64(\$personalAccessToken))

4 \$url = "https://app.vssps.visualstudio.com/_apis/accounts?api-version=7.1"

Send a Message

OpenAI Command-Line Chat

power of OpenAI's ChatGPT
your inputs and outputs are

--pattern summary

2025, potentially

AI with near-infinite
ing of long texts w

ind comprehensive know
ndows for complex p
ple interactions ar
tself, is predicted
s, as shown by ben
s in agent-based te
according to indus
implementation of

What is QakBot, and how can it be
exploited through the Follina
vulnerability?

AI Tim the Threat Intel Analyst

QakBot, also known as QBot or QAKBot, is a well-known threat actor group that has been active since 2018. They are primarily associated with delivering ransomware such as Emotet, Trickbot, and Ryuk. The group often uses spear-phishing emails to distribute malware, where they compromise email accounts and send malicious attachments or links in response to existing threads, aiming for a high rate of successful execution.

The Follina vulnerability (CVE-2022-30190) is related to the Windows msdt protocol handler, which allows attackers to execute arbitrary commands with elevated privileges.

Send a Message

LLMs can make mistakes. Verify important information.

pydscpts (workspace)

24 OPENAI_API_KEY = os.environ.get('OPENAI_API_KEY')

25 API_ENDPOINT = 'http://192.168.0.143:51443/api/chat/completions'

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285

286

287

288

289

290

291

292

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

588

589

590

591

592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

659

660

661

662

663

664

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702

703

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756

757

758

759

760

761

762

763

764

765

766

767

768

769

770

771

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810

811

812

813

814

815

816

817

818

819

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

Portable Consumption

```
christian@ctail-gce22007:~  
christian@ctail-gce22007:~  
$ fabric -y "https://youtu.be/g1JSZwvtEak/?feature=shared" --stream --pattern summarize extract_knowledge  
Here's a summary of the content in the requested format:  
  
ONE SENTENCE SUMMARY:  
Microsoft AI plans to release AI models with near-infinite memory in 2025, potentially revolutionizing AI capabilities and human-AI interactions.  
  
MAIN POINTS:  
1. Microsoft AI's head, Mustafa Suleyman, anticipates transformative AI with near-infinite memory by 2025.  
2. Google Research proposed "infinite attention" for efficient handling of long texts without excessive memory use.  
3. Infinite context windows could enable long-term AI relationships and comprehensive knowledge processing.  
4. Eric Schmidt discusses the potential of infinitely long context windows for complex problem-solving.  
5. Memory improvements will allow AI to maintain context across multiple interactions and sessions.  
6. Recursive self-improvement, where AI creates smarter versions of itself, is predicted before 2030.  
7. AI agents are still struggling with reliability in real-world tasks, as shown by benchmark tests.  
8. Current AI models show degrading performance over multiple attempts in agent-based tasks.  
9. Reliable autonomous agents may not be available until around 2026, according to industry experts.  
10. Significant improvements in error rates are needed for practical implementation of AI agents.  
  
TAKEAWAYS:  
1. Near-infinite memory in AI could dramatically enhance its capabilities and user interactions.  
2. Improved context retention may lead to more personalized and efficient AI experiences.  
3. Reliability remains a major challenge for AI agents in real-world applications.  
4. The development of AI is progressing rapidly, with significant advancements expected by 2025-2030.  
5. Successful implementation of AI agents requires substantial reduction in error rates for complex tasks.  
christian@ctail-gce22007:~  
$
```

15:14 40%

llama3.2:3b

Summarize what Chris is saying. I'll send you the transcript.

llama3.2:3b
I'm ready when you are. Please go ahead and share the transcript of the conversation with "Chris". I'll provide a concise summary.

One of these tools that I found was very cool from a Owasp is called owasp pyTM. This is open source from the wasp foundation and basically is a way of scripting. Writing code to generate my dfd. And I was like, you know, I wonder if I can use it like this. Unfortunately, the page is going up. But yeah, so I just said check dbt give me a cloud formation template that creates an EC2

+ Send a Message

LLMs can make mistakes. Verify important information.

Documentation

Mistral 7B - 33k (Laptop)

125 words per page - 264 pages

Phi-3 Medium 14B - 128k (Desktop)

300 words per page - 426 pages

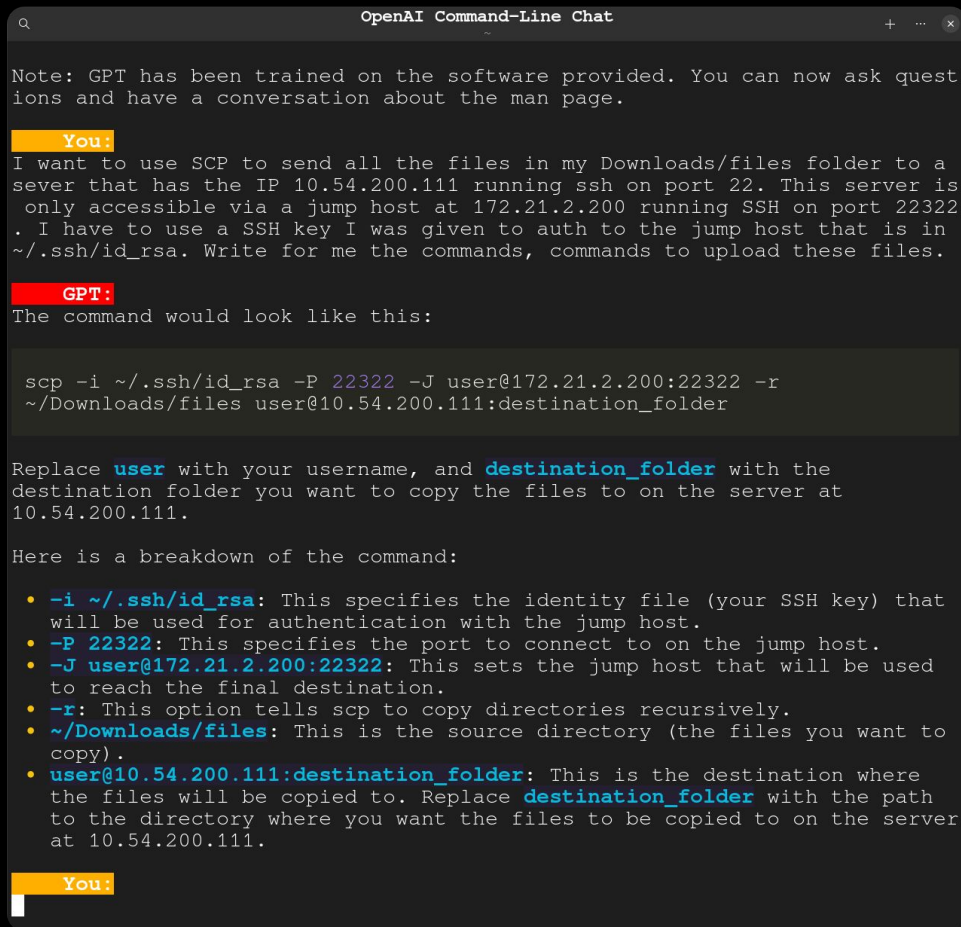
GPT-3.5 - 2k

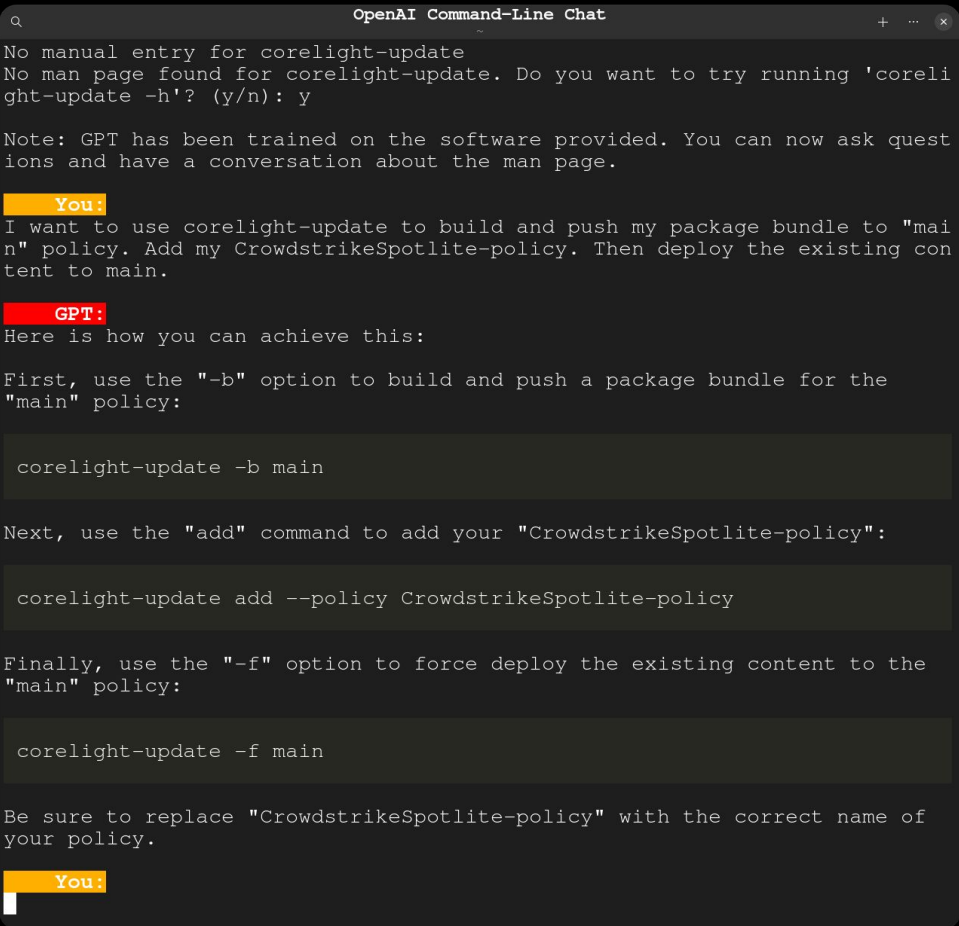
300 words per page - 5 pages

Hey, AI, add comments to my code.

Embedding Model

Embedding models are a type of machine learning model that learn to represent complex data, such as text or images as a number to calculate the similarity via assessing distance between data points.





AI Stacey the Scripter

```
110  
111     },  
112     animation: {  
113         duration: 1000  
114     },  
115     responsive: true,  
116     maintainAspectRatio: false  
117   }  
118 };  
119  
120 const ctx = document.getElementById('modelChart').getContext('2d');  
121 const myChart = new Chart(ctx, config);  
122  
123 // Set canvas height based on number of models  
124 document.querySelector('.chart-container').style.height = '100px';  
125 </script>  
126 </body>  
127 </html>
```

And the Python Flask server remains the same:

```
python  
1 from flask import Flask, render_template  
2  
3 app = Flask(__name__)  
4  
5 @app.route('/')  
6 def index():  
7     return render_template('index.html')  
8  
9 if __name__ == '__main__':  
10     app.run(debug=True)
```

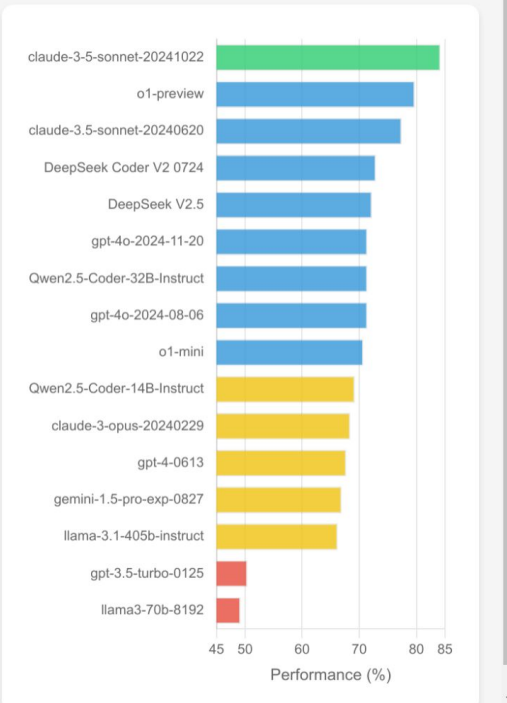
The improvements in this version include:

Run Code, Web Search

+ Send a Message

LLMs can make mistakes. Verify important information.

AI Model Performance Comparison



< Version 2 of 2 >

Copy

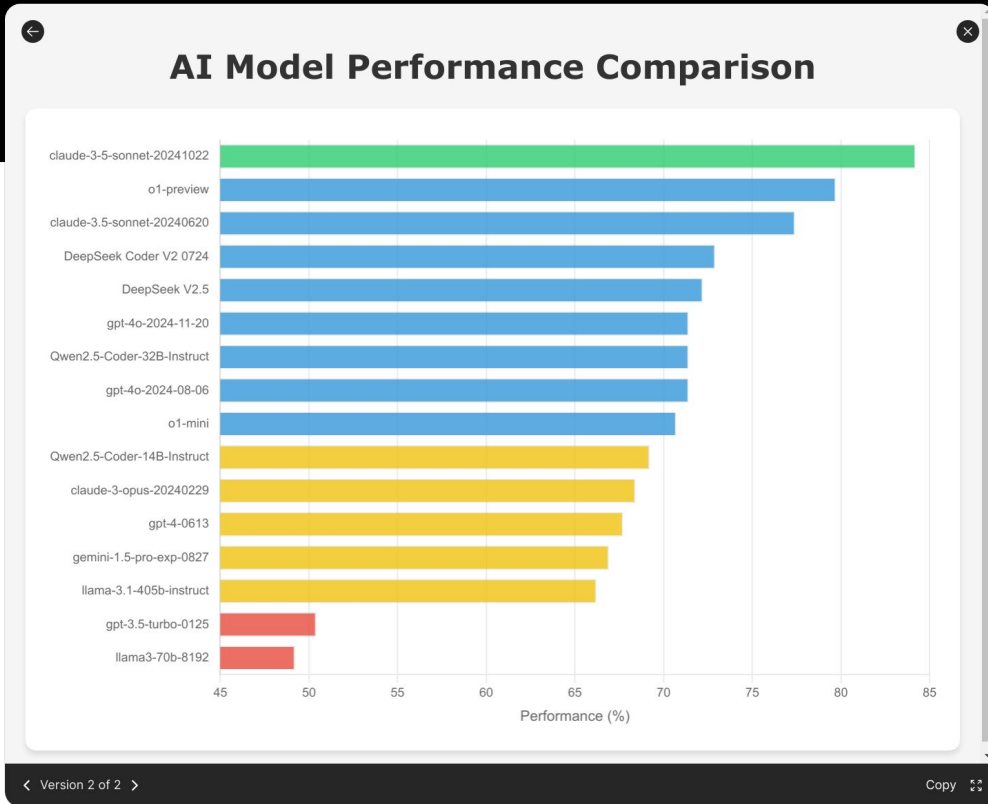


Stacey can write and present simple web apps, write in numerous languages, run python from chat, correct bugs, remediate vulns, and read repos.

Aider Coding Model

Stacey's Graph

[Aider's code editing benchmark](#) asks the LLM to edit python source files to complete 133 small coding exercises from Exercism. This measures the LLM's coding ability, and whether it can write new code that integrates into existing code. The model also has to successfully apply all its changes to the source file without human intervention.



But doesn't Copilot Write
Vulnerable Code?

Yep! But it can also fix it!

The screenshot shows a VS Code workspace with the following components:

- Editor (pom.xml):** Displays XML content for a Maven project. The `<parent>` section is highlighted, showing a dependency on `org.springframework.boot:spring-boot-starter-parent` with version `2.6.7`. The `<licenses>` section is also visible, listing the GNU Lesser General Public License.
- Terminal:** Shows a shell session with the following commands and output:

```
christian@ctail-gce22007: ~/github/clamav
master ✓ $ tree | xclip
christian@ctail-gce22007: ~/github/clamav
christian@ctail-gce22007: ~/github/clamav
master ✓ $
```
- Copilot Chat:** A chat window on the right side of the editor. It contains the following text:

The vulnerability you've mentioned is introduced through the `org.springframework.boot:spring-boot-starter-web` 2.6.7 dependency in your project, which transitively includes `org.springframework:spring-webmvc` 5.3.19. The fix for this vulnerability involves upgrading to a newer version of `spring-boot-starter-web`, specifically version 2.6.15.

To remediate the issue, you need to update the `pom.xml` file in your project that is located in the root directory of your project. Here's how you can do it:

 1. Open the `pom.xml` file.
 2. Locate the dependency for `org.springframework.boot:spring-boot-starter-web`.
 3. Change its version from 2.6.7 to 2.6.15.

Here is a snippet of what the updated section in your `pom.xml` might look like:

```
<dependency>
  <groupId>org.springframework.boot</groupId>
  <artifactId>spring-boot-starter-web</artifactId>
  <version>2.6.15</version>
</dependency>
```

After updating the version, save the changes and rebuild your project to ensure that the updated dependency is included.

If you are using a build tool other than Maven (e.g., Gradle), you will need to update the corresponding configuration file (`build.gradle` for Gradle) in a similar manner.

Here is the corrected content of your `pom.xml` file:

```
<?xml version="1.0" encoding="UTF-8"?>
<project xmlns="http://maven.apache.org/POM/4.0.0"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://maven.apache.org/POM/4.0.0 http://maven.apache.org/maven-v4_0_0.xsd">

  <modelVersion>4.0.0</modelVersion>

  <parent>
```

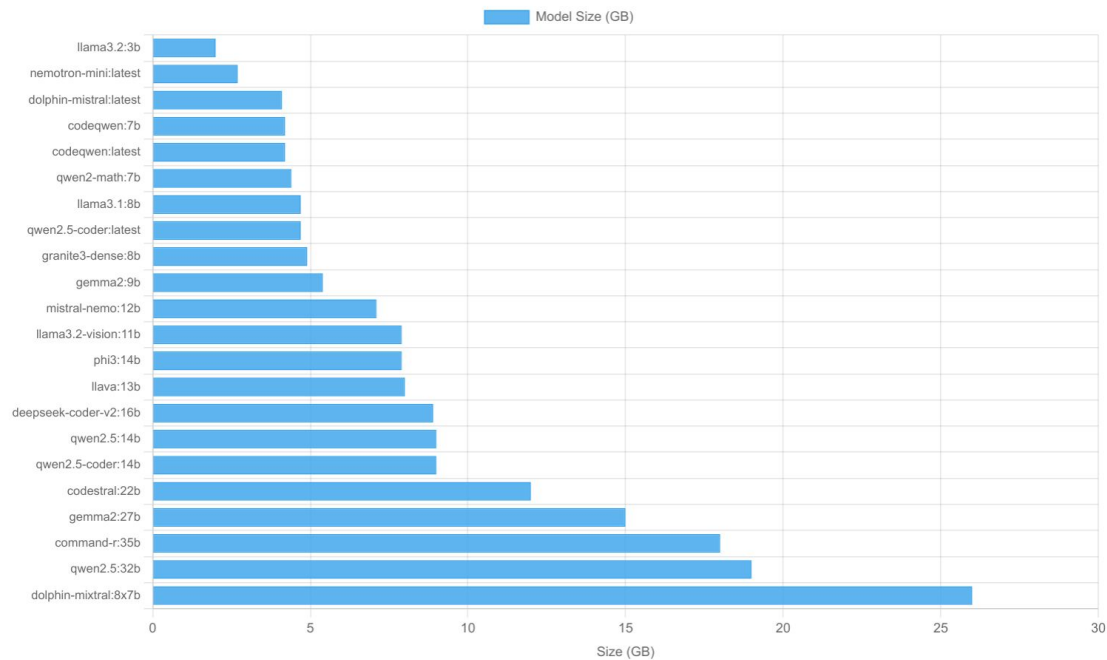
03

Models

Models

And their sizes

AI Models Size Comparison



Knowledge and Model Size

Larger Models

- Generally more capable
- Expensive to train
- Generally Lack of Specialized Models
- Comes with greater world knowledge
- Greater Zero Shot Logic and Reasoning
- Slower
- Increased hardware requirements

Smaller Models

- Generally less capable
- Cheap to train
- Specialized Models
- Less world knowledge (besides specialization)
- Weaker Zero Shot Logic and Reasoning
- Faster
- Reduced hardware requirements

Quantization to reduce model size at precision cost.

Augment LLM

Fine Tuning

Tune to modify or refine model behavior, give model new skill or capability. Offers the capability to change how the model acts.

Syntax learning, writing style, formatting, are all useful examples of modifications that would benefit from fine tuning.

RAG

Use RAG to modify models knowledge of facts. Provides more reliable consistent fact recall by using AI to call relevant information.

Searching the Internet, knowledge base, or your SIEM are all useful examples of modifications that would benefit from fine tuning.

Machine Learning Profiles/GPTs/Modelfiel - Old Agents

GIST

AI Madison the Cybersecurity Manager

Model qwen2.5:32b

A cybersecurity manager responsible for cybersecurity strategy and operations with a focus on security policies, risk management, team leadership, and strategic planning within the cybersecurity domain.

Profile: Prompt + Knowledge + Tools + Actions + Capabilities



≡ Models Knowledge Prompts Tools

Tools
To select toolkits here, add them to the "Tools" workspace first.
☒ Web Search ☐ Run Code ☐ Web Scrape

Filters
To select filters here, add them to the "Functions" workspace first.
☐ MoEA

Actions
To select actions here, add them to the "Functions" workspace first.
☐ Visualize Data

Capabilities
☐ Vision ☐ Usage ☒ Citations

04

Challenges

Challenges

Use Case Issue

Incorrect application of this technology because it is “easy”. Many traditional tools are still the best for their job.

Example: Performing anomaly detection, or other big data science function.

Model Weakness

Use statistical correlation to reflect and leverage the patterns and purpose of language - partially inferring logical construct.

Real-world situations involve incredible nuances. Prompts are often the size of a tweet.

Knowledge & Context Limits

Models are trained up to a certain point of time. RAG can accommodate this, but context is limited.

Note: Although context size may be large, models can lose track of information relevance.

Reliability of Probability

We work with risk and probability all the time.

At the end of the day, it is a guess. Just like our modern machine learning.

What is our “false” tolerance?



Model Security

Deserialization Threats

Untrusted data/code reconstructs objects, leading to exploitation.

Malicious actors inject harmful code during deserialization.

Crucial for securing data integrity and preventing unauthorized code execution.

[models-scan](#)

Architectural Backdoors

Parallel path from model input to output, manipulated by attackers.

Model behaves normally with non-malicious inputs, unexpected behavior with triggers.

Attacker controls when backdoored model gives incorrect output.

Runtime Threats

Untrusted data/code exploits vulnerabilities during inference or execution.

Malicious actors inject harmful code, gaining unauthorized access or manipulating systems.

Crucial for securing data integrity and preventing unauthorized code execution.

05

Tooling

FOSS Stack

ZTNA + Firewall Decrypted

OpenWeb UI | Obsidian

OpenAI API | Auth Proxy

Ollama | vLLM | Auth Proxy

STT:Whisper | TTS:SpeechT5

RAG: nomic-embed-text

Why Ollama?

Because it is easy.

```
ollama run llama3.2:3b --verbose
christian@michelangelo:~
$ ollama pull llama3.2:3b
pulling manifest
pulling dde5aa3fc5ff... 100% 2.0 GB
pulling 966de95ca8a6... 100% 1.4 KB
pulling fcc5a6bec9da... 100% 7.7 KB
pulling a70ff7e570d9... 100% 6.0 KB
pulling 56bb8bd477a5... 100% 96 B
pulling 34bb5ab01051... 100% 561 B
verifying sha256 digest
writing manifest
success
christian@michelangelo:~
$ ollama run llama3.2:3b --verbose
>>> Hello there!
It's nice to meet you. Is there something I can help you with
or would you like to chat?

total duration:      196.066373ms
load duration:       30.393ms
prompt eval count:   28 token(s)
prompt eval duration: 23ms
prompt eval rate:    1217.39 tokens/s
eval count:          23 token(s)
eval duration:       142ms
eval rate:           161.97 tokens/s
>>> Send a message (? for help)
```


“GPU Poor” for Testing

XPS 15 2020

Old Corporate Laptop

Intel i7-10750H

NVIDIA GTX 1650 Ti Mobile

4GB VRAM

32 GB RAM

~7b at reading speed (5t/s)

ThinkStation P510

Old Donated Workstation

Intel Xeon E5-2609 v4

NVIDIA GTX 1080

8GB VRAM

96GB RAM

~9b at reading speed (12t/s)

Gaming Desktop

~\$1k Desktop

12th Gen Intel i7-12700F

NVIDIA GeForce RTX 3080

10GB VRAM

32GB RAM

32b at reading speed (6ts)

Recommendation: High VRAM GPU (3090 or 4090)

2x3090 = 48GB VRAM = llama3.1:70b | reflection:70b | nvidia nemotron:70b

The screenshot displays the Obsidian application interface. On the left, a sidebar contains various icons for file management and navigation. The main area is split into two panes. The left pane shows a 'Graph view' of a knowledge graph with numerous nodes (colored dots) and edges (lines) connecting them. A central node is labeled 'COZY BEAR'. The right pane is a chat window titled 'What would you like to talk about with your notes?'. It contains a user message: 'Tell me about the Adversary Cozy Bear (APT 29). What have they been up to recently?'. Below this is a response from the AI model, which includes a section header 'Cozy Bear (APT 29) Recent Activities' and two paragraphs of text. The first paragraph describes Cozy Bear as a Russian threat actor and mentions a spear-phishing campaign. The second paragraph describes the content of the spear-phishing emails. The chat window has a status bar at the bottom showing '10 files 0 min read 55s'.

Graph view

What would you like to talk about with your notes?

Tell me about the Adversary Cozy Bear (APT 29). What have they been up to recently?

Cozy Bear (APT 29) Recent Activities

Cozy Bear, also known as UNC2452 or APT29, is a Russian threat actor that has been actively engaged in malicious activities. Recently, they have been sending targeted spear-phishing emails to thousands of individuals in over 100 organizations since October 22, 2024 [actra/programs/vsrt-exchange/2024/2024-11-November-vsrt](#).

These spear-phishing emails contain a signed Remote Desktop Protocol (RDP) configuration file that connects to an actor-controlled server, potentially exposing sensitive information and allowing the installation of malware. The campaign aims to collect intelligence, indicating a continued threat from Cozy Bear in the recent months [actra/programs/vsrt-exchange/2024/2024-11-November-vsrt](#).

It's essential for organizations to remain vigilant and take necessary precautions to protect themselves against such targeted attacks.

Chat with your Smart Second Brain...

10 files 0 min read 55s

Talk with your knowledge management. Embed files with **nomic-embed-text** and allow statistics and probability to find the relevant files for your to have conversations with your data using a light model like **llama3.2:3b**.

Obsidian + Second Brain + Ollama

Use Continue.dev to connect to various third party APIs for inference, OpenAI compatible AIs, or Ollama in a CoPilot like experience.

Continue.dev + OpenWebUI

The screenshot displays the Continue.dev IDE interface. The main editor shows a `pom.xml` file for a project named `clamav-rest`. The file contains XML metadata, parent and group information, and a license section. The chat window on the right provides context-aware suggestions for updating the `spring-boot-starter-web` dependency to version 2.6.15 to address a vulnerability. The terminal window at the bottom shows a user running `tree` and `xclop` commands in a shell.

```
<?xml version="1.0" encoding="UTF-8"?>
<project xmlns="http://maven.apache.org/POM/4.0.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://maven.apache.org/POM/4.0.0 http://maven.apache.org/maven-v4_0_0.xsd">
  <modelVersion>4.0.0</modelVersion>
  <parent>
    <groupId>org.springframework.boot</groupId>
    <artifactId>spring-boot-starter-parent</artifactId>
    <version>2.6.7</version>
  </parent>
  <groupId>fi.solita.clamav</groupId>
  <artifactId>clamav-rest</artifactId>
  <version>1.0.2</version>
  <packaging>jar</packaging>
  <name>Simple ClamAV proxy</name>
  <description>Simple REST proxy to put between ClamAV and a web server</description>
  <url>https://github.com/solita/clamav-rest</url>
  <licenses>
    <license>
      <name>GNU LESSER GENERAL PUBLIC LICENSE</name>
      <url>http://www.gnu.org/licenses/lgpl-3.0.html</url>
    </license>
  </licenses>
  <developers>
    <developer>
      <name>Antti Virtanen</name>
      <email>antti.virtanen@solita.fi</email>
    </developer>
  </developers>
</project>
```

The vulnerability you've mentioned is introduced through the `org.springframework.boot:spring-boot-starter-web` 2.6.7 dependency in your project, which transitively includes `org.springframework:spring-webmvc` 5.3.19. The fix for this vulnerability involves upgrading to a newer version of `spring-boot-starter-web`, specifically version 2.6.15.

To remediate the issue, you need to update the `pom.xml` file in your project that is located in the root directory of your project. Here's how you can do it:

1. Open the `pom.xml` file.
2. Locate the dependency for `org.springframework.boot:spring-boot-starter-web`.
3. Change its version from 2.6.7 to 2.6.15.

Here is a snippet of what the updated section in your `pom.xml` might look like:

```
<dependency>
  <groupId>org.springframework.boot</groupId>
  <artifactId>spring-boot-starter-web</artifactId>
  <version>2.6.15</version>
</dependency>
```

After updating the version, save the changes and rebuild your project to ensure that the updated dependency is included.

If you are using a build tool other than Maven (e.g., Gradle), you will need to update the corresponding configuration file (`build.gradle` for Gradle) in a similar manner.

Here is the corrected content of your `pom.xml` file:

```
<?xml version="1.0" encoding="UTF-8"?>
<project xmlns="http://maven.apache.org/POM/4.0.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://maven.apache.org/POM/4.0.0 http://maven.apache.org/maven-v4_0_0.xsd">
  <modelVersion>4.0.0</modelVersion>
  <parent>
    <groupId>org.springframework.boot</groupId>
    <artifactId>spring-boot-starter-parent</artifactId>
    <version>2.6.15</version>
  </parent>
  <groupId>fi.solita.clamav</groupId>
  <artifactId>clamav-rest</artifactId>
  <version>1.0.2</version>
  <packaging>jar</packaging>
  <name>Simple ClamAV proxy</name>
  <description>Simple REST proxy to put between ClamAV and a web server</description>
  <url>https://github.com/solita/clamav-rest</url>
  <licenses>
    <license>
      <name>GNU LESSER GENERAL PUBLIC LICENSE</name>
      <url>http://www.gnu.org/licenses/lgpl-3.0.html</url>
    </license>
  </licenses>
  <developers>
    <developer>
      <name>Antti Virtanen</name>
      <email>antti.virtanen@solita.fi</email>
    </developer>
  </developers>
</project>
```

main* 0 0 0 0 8 mins -- NORMAL -- Ln 21, Col 12 Spaces: 2 UTF-8 LF XML ✓ Continue

06

Frontier

AI Ceiling?

Why are community models catching up to the big players?

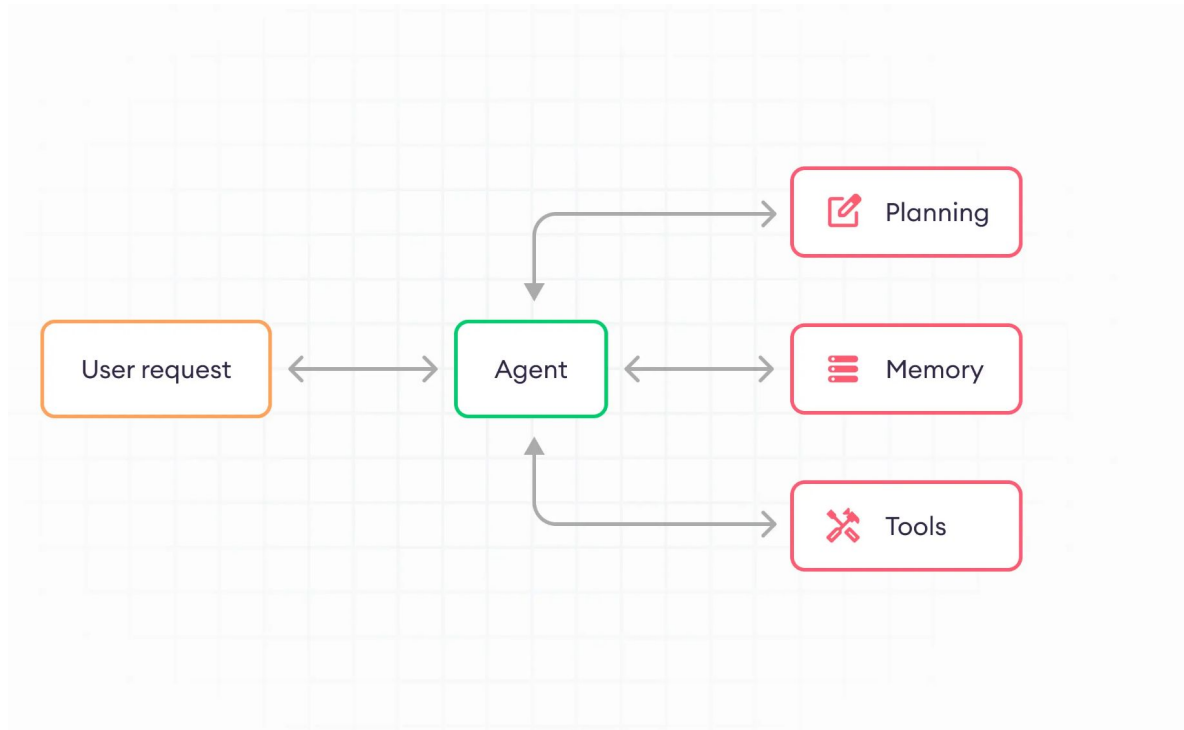
LLM Agency

Allowing LLMs to determine when to use which tools. Proving LLMs with the capabilities to “do” things.

Having LLMs think through a process and assess its results. Reconsider if the results are unexpected.

Multishot + Agency

- Plan
- Reflect
- Search the internet
- Scrape a web pack
- Run code



Accelerating Improvements

RAG Embedding

Reduce storage capacity requirements

Fasters

Reranking

Implication: arger datasets can fuel RAG effectively

Quantization

More strategic quantization that preserves knowledge and removes noise.

Implication: Smaller models can be more effective.

Fine Tuning

Better data

Tuning Efficiency

Strategic tuning

Implication: Smaller models can be more effective.

Hardware Scale

Hardware acceleration designed for AI tasks

Implication: Costs of running AI will continue to shrink

THANK YOU

<https://christiant.io>

Taillon, Christian

