AUTOMATTIC

Red Tide Revert

David Newman
AI Systems
@darthhexx

1. Red Tide
2. Reverts
3. Rapid Iteration
4. Automation
5. Augmentation
6. Future State

What does an ocean tide have to do with a revert...


Digital Retail Signage
TECH X


Retail Media Networks
Power tools screen
R 650.00
R 53,625.00
65,204
Customers reached
R 643,500.00


Digital Menu Boards


Large-format LED

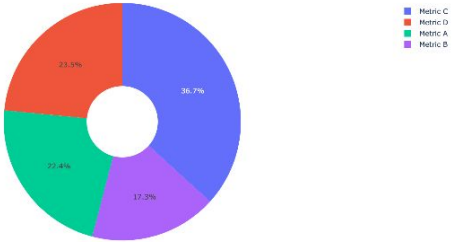

Interactive Touchscreen

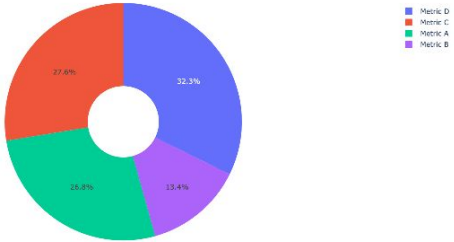

QSR Drive Thru Solutions
KFC BUCKETS
FAMILY TREAT
2 CAN DINE
ADD TO YOUR MEAL

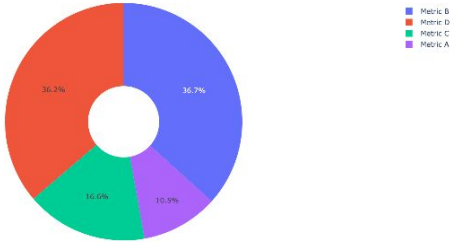
Load-shedding

****

Metrics Dashboard

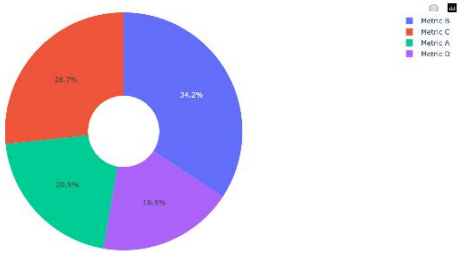


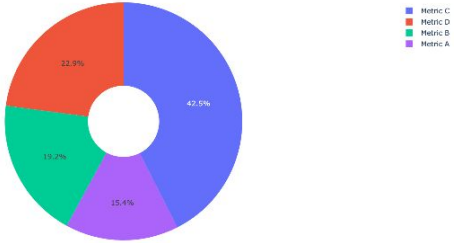


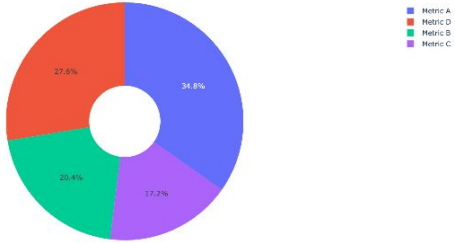


Name	Value	Status
Item 1	446	Online
Item 2	665	Online
Item 3	192	Online
Item 4	670	Online
Item 5	593	Online
Item 6	375	Online
Item 7	638	Online
Item 8	829	Online
Item 9	437	Online
Item 10	682	Online
Item 11	835	Online
Item 12	816	Online
Item 13	455	Online
Item 14	439	Online
Item 15	758	Online
Item 16	107	Online
Item 17	408	Online
Item 18	199	Online
Item 19	317	Online
Item 20	349	Online
Item 21	320	Online
Item 22	418	Online
Item 23	448	Online
Item 24	589	Online
Item 25	425	Online
Item 26	504	Online
Item 27	785	Online
Item 28	967	Online
Item 29	662	Online
Item 30	449	Online
Item 31	834	Online
Item 32	504	Online
Item 33	779	Online
Item 34	875	Online

Metrics Dashboard

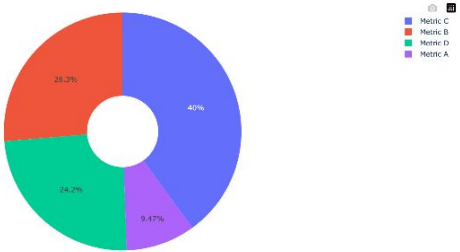


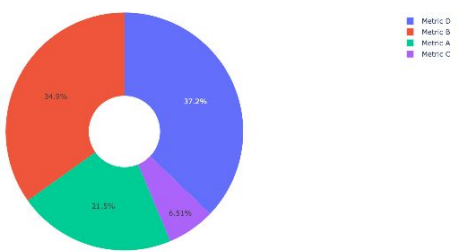


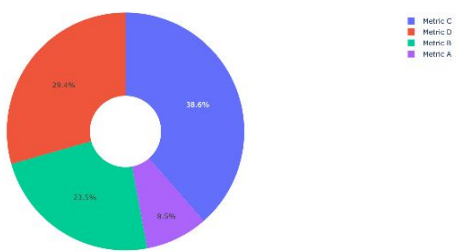


Name	Value	Status
Item 1	914	Offline
Item 2	850	Offline
Item 3	939	Offline
Item 4	959	Offline
Item 5	548	Offline
Item 6	421	Online
Item 7	208	Online
Item 8	439	Online
Item 9	613	Online
Item 10	544	Online
Item 11	885	Online
Item 12	377	Online
Item 13	561	Online
Item 14	385	Online
Item 15	408	Online
Item 16	268	Online
Item 17	284	Online
Item 18	322	Online
Item 19	682	Online
Item 20	225	Online
Item 21	859	Online
Item 22	260	Online
Item 23	816	Online
Item 24	312	Online
Item 25	308	Online
Item 26	887	Online
Item 27	588	Online
Item 28	349	Online
Item 29	321	Online
Item 30	689	Online
Item 31	246	Online
Item 32	264	Online
Item 33	556	Online
Item 34	230	Online

Metrics Dashboard

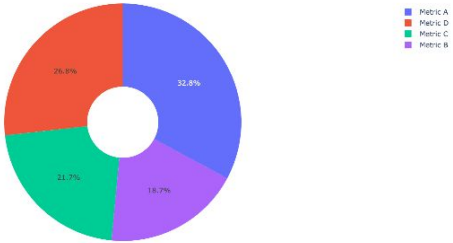


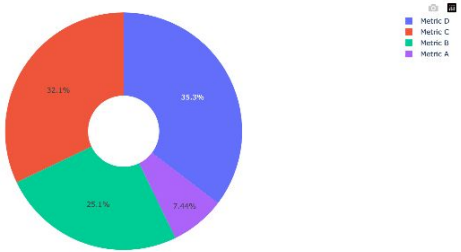


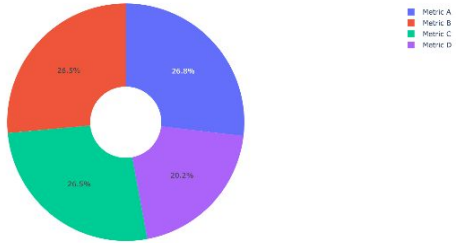


Name	Value	Status
Item 1	597	Offline
Item 2	278	Offline
Item 3	940	Offline
Item 4	980	Offline
Item 5	498	Offline
Item 6	183	Offline
Item 7	614	Offline
Item 8	112	Offline
Item 9	403	Offline
Item 10	528	Online
Item 11	447	Online
Item 12	477	Online
Item 13	408	Online
Item 14	927	Online
Item 15	535	Online
Item 16	858	Online
Item 17	620	Online
Item 18	578	Online
Item 19	195	Online
Item 20	484	Online
Item 21	967	Online
Item 22	331	Online
Item 23	357	Online
Item 24	578	Online
Item 25	974	Online
Item 26	854	Online
Item 27	184	Online
Item 28	716	Online
Item 29	823	Online
Item 30	431	Online
Item 31	396	Online
Item 32	293	Online
Item 33	796	Online
Item 34	449	Online

Metrics Dashboard

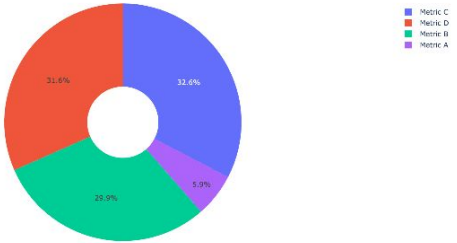


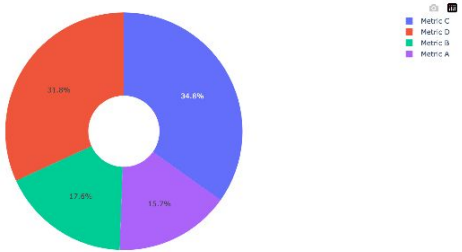


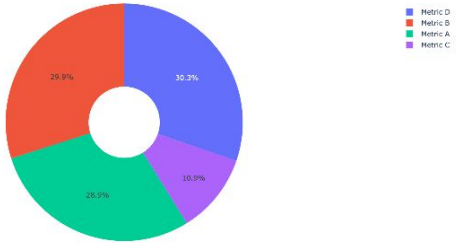


Name	Value	Status
Item 1	865	Offline
Item 2	172	Offline
Item 3	808	Offline
Item 4	904	Offline
Item 5	605	Offline
Item 6	532	Offline
Item 7	354	Offline
Item 8	900	Offline
Item 9	901	Offline
Item 10	250	Offline
Item 11	428	Offline
Item 12	541	Offline
Item 13	339	Offline
Item 14	927	Offline
Item 15	492	Offline
Item 16	939	Online
Item 17	151	Online
Item 18	481	Online
Item 19	330	Online
Item 20	762	Online
Item 21	379	Online
Item 22	828	Online
Item 23	181	Online
Item 24	167	Online
Item 25	337	Online
Item 26	251	Online
Item 27	429	Online
Item 28	215	Online
Item 29	175	Online
Item 30	630	Online
Item 31	524	Online
Item 32	125	Online
Item 33	166	Online
Item 34	242	Online

Metrics Dashboard







Name	Value	Status
Item 1	427	Offline
Item 2	363	Offline
Item 3	360	Offline
Item 4	402	Offline
Item 5	166	Offline
Item 6	102	Offline
Item 7	382	Offline
Item 8	377	Offline
Item 9	806	Offline
Item 10	695	Offline
Item 11	149	Offline
Item 12	524	Offline
Item 13	701	Offline
Item 14	143	Offline
Item 15	731	Offline
Item 16	420	Offline
Item 17	546	Offline
Item 18	474	Offline
Item 19	116	Offline
Item 20	288	Offline
Item 21	635	Offline
Item 22	124	Offline
Item 23	161	Offline
Item 24	606	Offline
Item 25	280	Offline
Item 26	772	Offline
Item 27	411	Offline
Item 28	463	Offline
Item 29	577	Offline
Item 30	958	Offline
Item 31	292	Online
Item 32	373	Online
Item 33	648	Online
Item 34	416	Online

1. Red Tide
2. Reverts
3. Rapid Iteration
4. Automation
5. Augmentation
6. Future State

What is a revert...

Nagios

PagerDuty

Grafana OnCall

dynatrace

splunk>

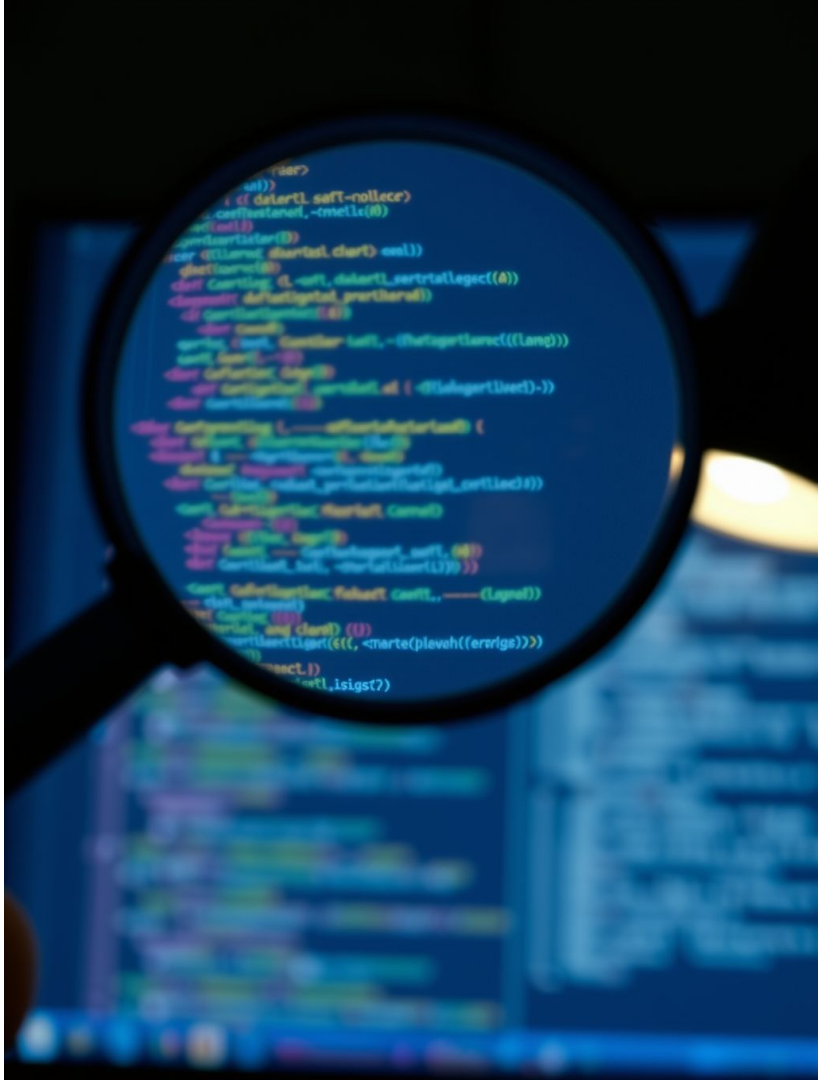
slack

new relic

DATADOG

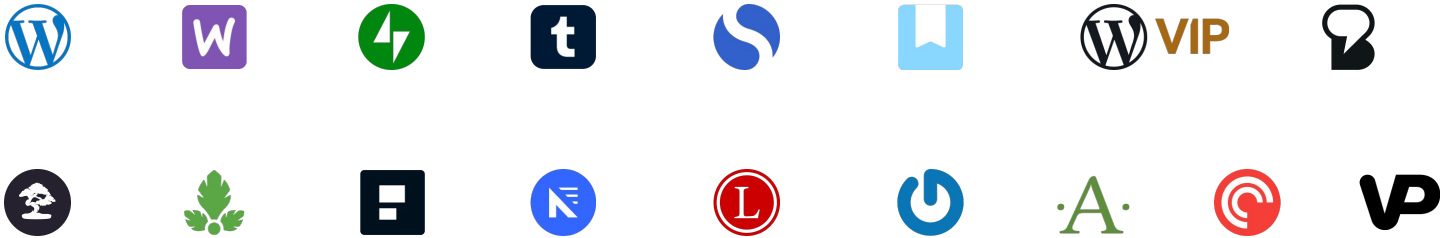
Alerts to Reverts

- Engineers investigate alerts
- Request patterns?
  - Block
- Previously unidentified 🪲?
  - Patch
- Recent change in production?
  - Revert the change (rollback)



1. Red Tide
2. Reverts
3. Rapid Iteration
4. Automation
5. Augmentation
6. Future State

AUTOMATTIC



WordPress.com

WE ITERATE

942

Deployments made this week

view graph

WE COMMUNICATE

328,449

Messages sent this week

view details

WE MAKE PEOPLE HAPPY

29,512

Support interactions this week

view graph

Challenges of Rapid Iteration

- Blue/Green deploys can’t scale
- Staged deploys, i.e. Canary → Low → Mid → High traffic, doesn’t apply
- Mono-repo “issues”
- Requires rapid rollback capability

1. Red Tide
2. Reverts
3. Rapid Iteration
4. Automation
5. Augmentation
6. Future State

What can we put in place to reduce TOIL and stress...

Work With Us, from anywhere

We are a truly distributed company, which means you can work from the places where you feel most inspired and productive. We believe in the power of flexibility and trust, empowering all “Automatticians” to find the perfect work-life balance while delivering the highest impact.


Beaufort Sea
Canada
United States
Mexico
Cuba
Venezuela
Colombia
Ecuador
Brazil
Bolivia
Paraguay
Argentina
Chile
Falkland Islands (Islas Malvinas)
Scotia Sea
North Pacific Ocean
South Pacific Ocean
North Atlantic Ocean
South Atlantic Ocean
Sweden
Finland
Russia
Belarus
Poland
Germany
France
Romania
Kazakhstan
Mongolia
Uzbekistan
Tajikistan
Turkey
Iraq
Iran
Afghanistan
Pakistan
China
Morocco
Algeria
Tunisia
Libya
Egypt
Saudi Arabia
Mauritania
Mali
Niger
Chad
Sudan
Yemen
Ethiopia
Nigeria
Angola
Zambia
Namibia
Zimbabwe
Madagascar
Tanzania
Democratic Republic of the Congo
India
Nepal
Laos
Indonesia
Papua New Guinea
Australia
Arabian Sea
Indian Ocean
Coral Sea
Bering Sea

AUTOMATTIC

1734

Automatticians [1]

92

Countries [1]

116

Languages spoken [1]

1. Data taken from live data on 10th October 2024 from automattic.com/about


Meetup


ATHENS

Meetup Projects

Iterate on an existing feature

Team goal related

High-impact on TOIL reduction

User facing feature

Dr. Fix It

- ELK stack
- WordPress-backed PHP service
- Slack Application

Dr. Fix It APP 20:53

🔴 **Warning:** around `2024-10-10 10:50:35 UTC` we got `51` of the following error (Kibana):

```
Uncaught Exception: Currencies must be the same (adding EUR to USD)
```

1️⃣ **Possible cause:** 582c85d214f53c3d4f3d404aef3d42c83c1ce213
Deployed `9 minutes` ago at `2024-10-10 10:37:06` in `wpcom-git` by `@[redacted]`

```
Embeds: Bump to 50%

Reviewers: #devops_team!

Differential Revision: https://[redacted]/D163538
```

2️⃣ **Possible cause:** db9b04e03f0e794bd6f222d9d71750ca0a105965
Deployed `4 minutes` ago at `2024-10-10 10:46:50` in `wpcom-git` by `@[redacted]`

```
Removing [redacted] @[redacted] from roundrobin pings

Summary:
For each form submission in agency-engagement-request-form on
https://[redacted].wordpress.com/[redacted]/ - a new
post is created with the submission data as content.
This DIFF removes [redacted] from the pingable people list, so it should only be
pinging two folks from now on.

Test Plan:
- Open https://[redacted].wordpress.com/[redacted]/
- Fill the form and submit
- Confirm that no errors appear and the form submission summary is shown
- CHeck that a new Post was created and that it has all the data submitted in
the form
- Please consider that testing this (even locally) will create a new post and
will tag someone (unless you change the username being tagged in the post)

Differential Revision: https://[redacted]/D163537
```

3️⃣ **Possible cause:** d5f350b11e5ad955f7ed91d7e101e265d664c399
Deployed `4 minutes` ago at `2024-10-10 10:48:15` in `wpcom-git` by `@[redacted]`

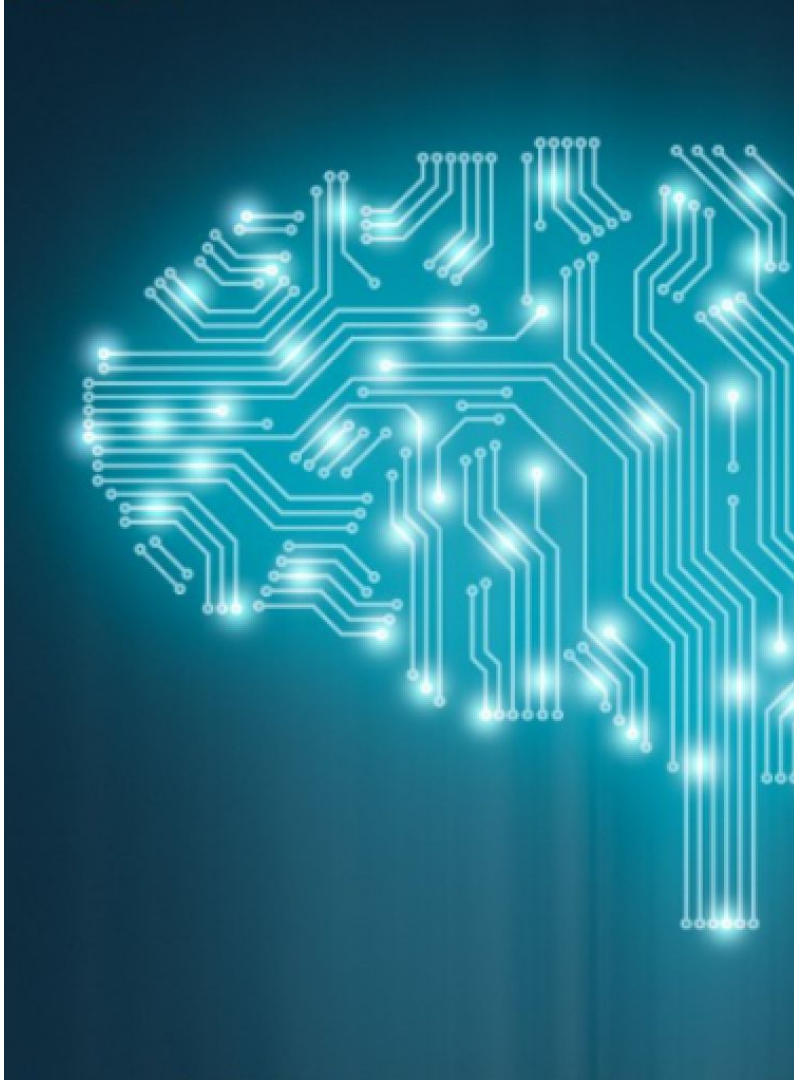
```
Revert "Embeds: Bump to 50%"

This reverts commit 582c85d214f53c3d4f3d404aef3d42c83c1ce213.
```

✅ 1

Dr. Fix It - Ideation

- Use AI to determine if a commit caused the stack trace we get in the logs
- Only display a filtered commit list to reduce on-call noise
- Command for engineers to easily revert
- Stretch goal: AI to execute the revert



1. Red Tide
2. Reverts
3. Rapid Iteration
4. Automation
5. Augmentation
6. Future State

Dr. Fix It - The Plan

- Gather range test cases
- Craft a ZeroShot prompt to determine if a commit → stack trace
- Use our internally hosted LLM services

Dr. Fix It - The Initial Test

- Crafted a simple direct prompt
- PoC service in our Ray cluster
- Prepared a tiny test set
- Quick test run before dinner...

2 out of 3.

That's not too bad, but why do different LLMs agree that the same commit *didn't* cause the stack trace?

Dr. Fix It - The PoC Test

- ZeroShot prompt
- Llama3.1 70B vLLM service
- Ray Serve python service for AI agent
- 14 cherry picked, and double-checked, test cases
- 14 negative tests

Best run was 26 out of 28.

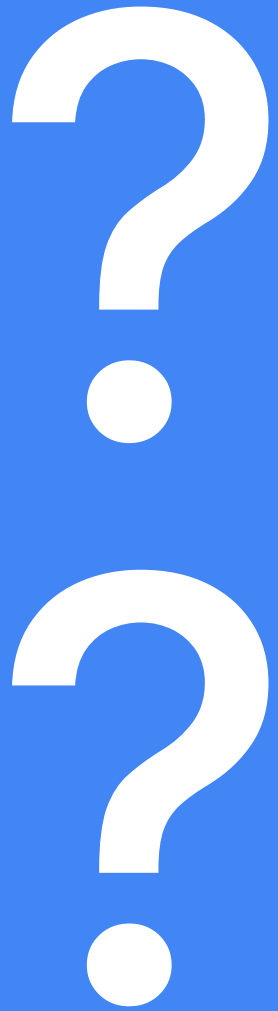
Wow, okay. All we need to do is add agentic reasoning, tool calling, and guardrails and we'll be all set to deploy.


HEY! GET BACK TO WORK!
Agents finding root cause!
OH. CARRY ON.

SRE CON®

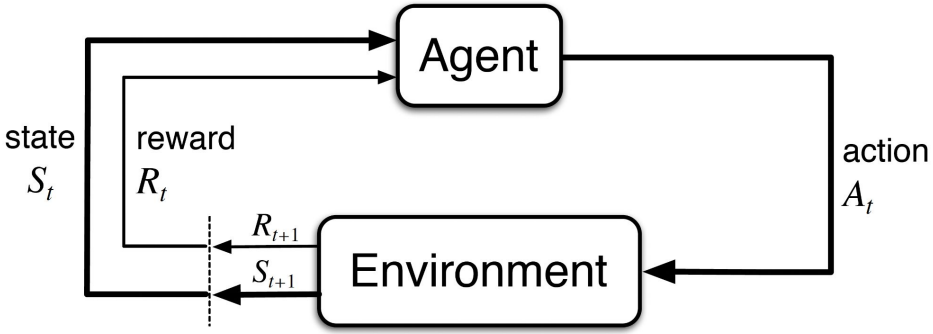
EUROPE
MIDDLE EAST
AFRICA

Now, about those two test cases...



AI Agent

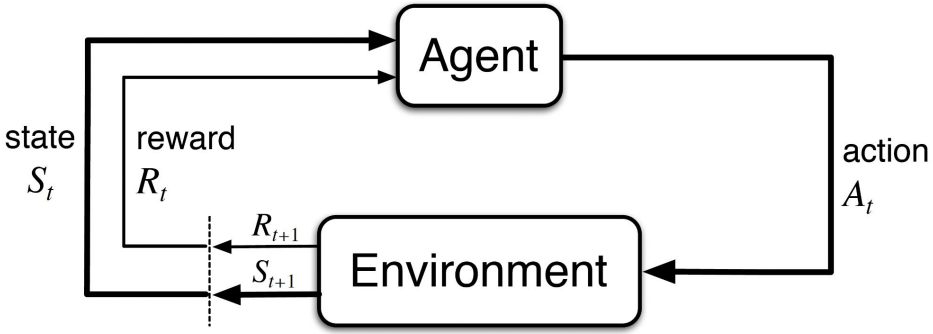
- Observe environment
- Agent evaluates
- Perform an action
- Environmental reward
- Loop until end of episode or forever



Reinforcement Learning: An Introduction. Sutton and Barto

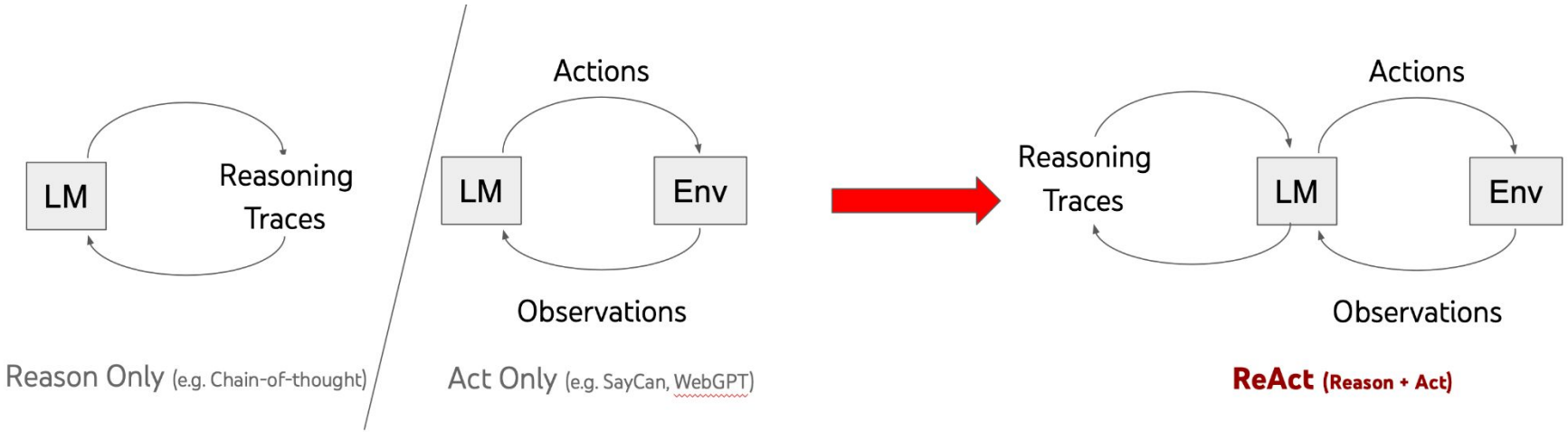
LLM Agent

- Receive an observation - **Prompt**
- Agent evaluation - **Reasoning**
- Perform an action - **Tools**
- Reward - depends on cognitive architecture
- Loop until final answer



Reinforcement Learning: An Introduction. Sutton and Barto

Reasoning & Tools → ReACT

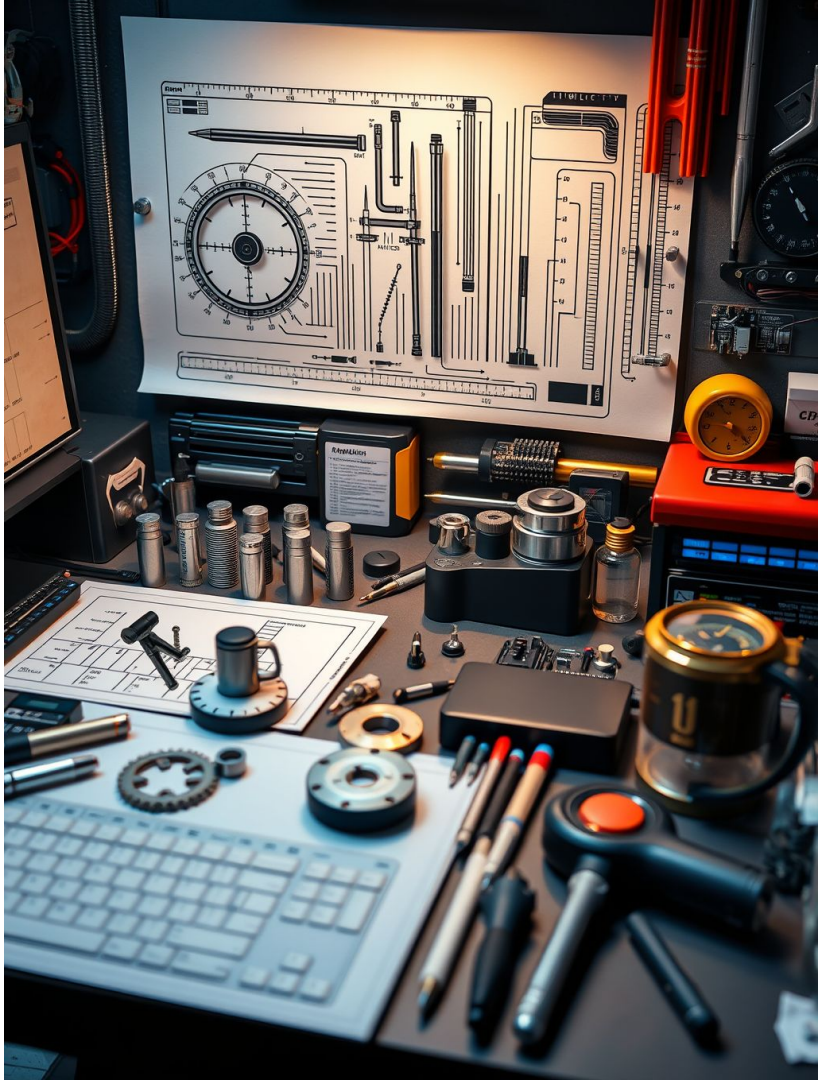


ReAct: Synergizing Reasoning and Acting in Language Models: arxiv.org/abs/2210.03629

Agent Tools

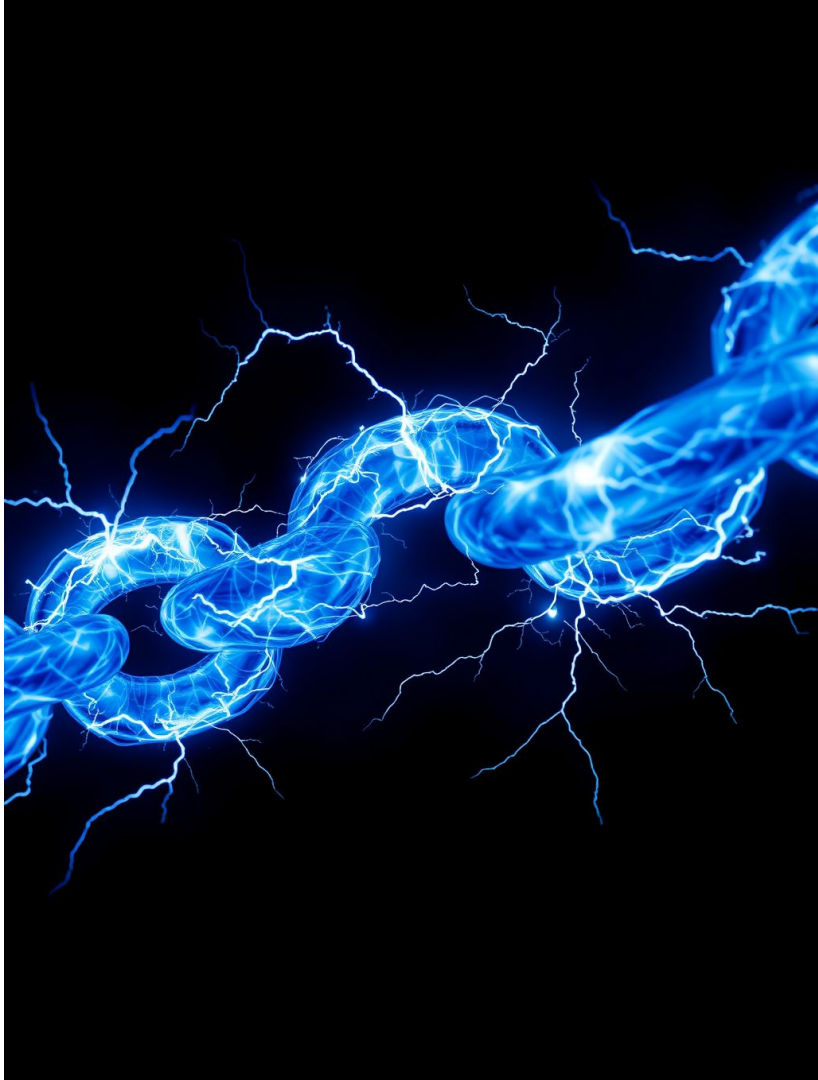
- Function definition
- Function source code
- Class function definition
- Class source code
- Git blame-based functions
- Reflection [1]

1. Reflexion: Language Agents with Verbal Reinforcement Learning. https://arxiv.org/abs/2303.11366



Chain of Thought

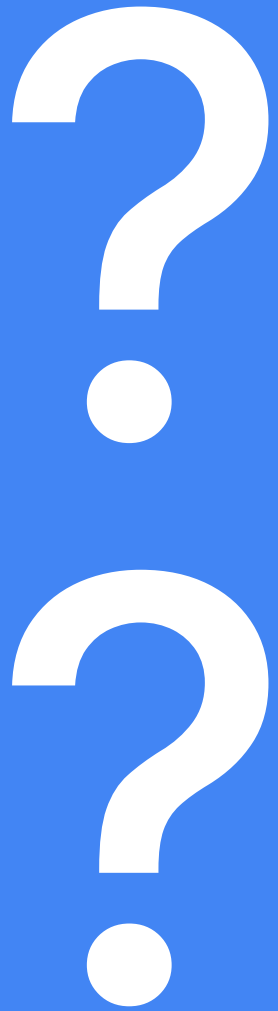
- Chain-of-Thought with self-consistency
- Zero-Shot Chain-of-Thought
- Automatic-Chain-of-Thought
- Program-of-Thoughts prompting
- Tree-of-Thoughts
- Graph-of-Thoughts
- Algorithm-of-Thoughts
- Skeleton-of-Thought
- Buffer-of-Thought
- Logic-of-Thought



Open-(Weight|Source) LLMs

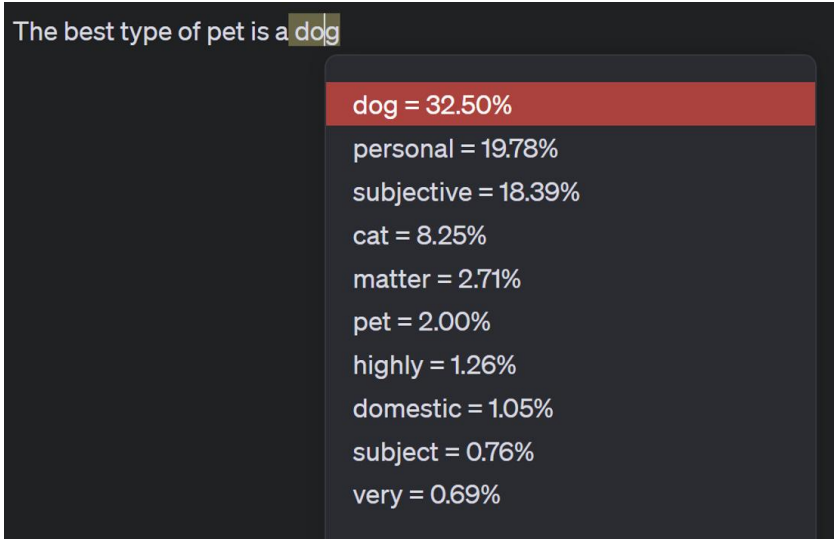
- DeepseekV2 Coder
- Granite Code
- Llama 3.1
- Phi 3.5
- Qwen 2.5 Coder
- StarCoder2
- ToolACE
- WaveCoder
- Yi-Coder

And those two test cases...

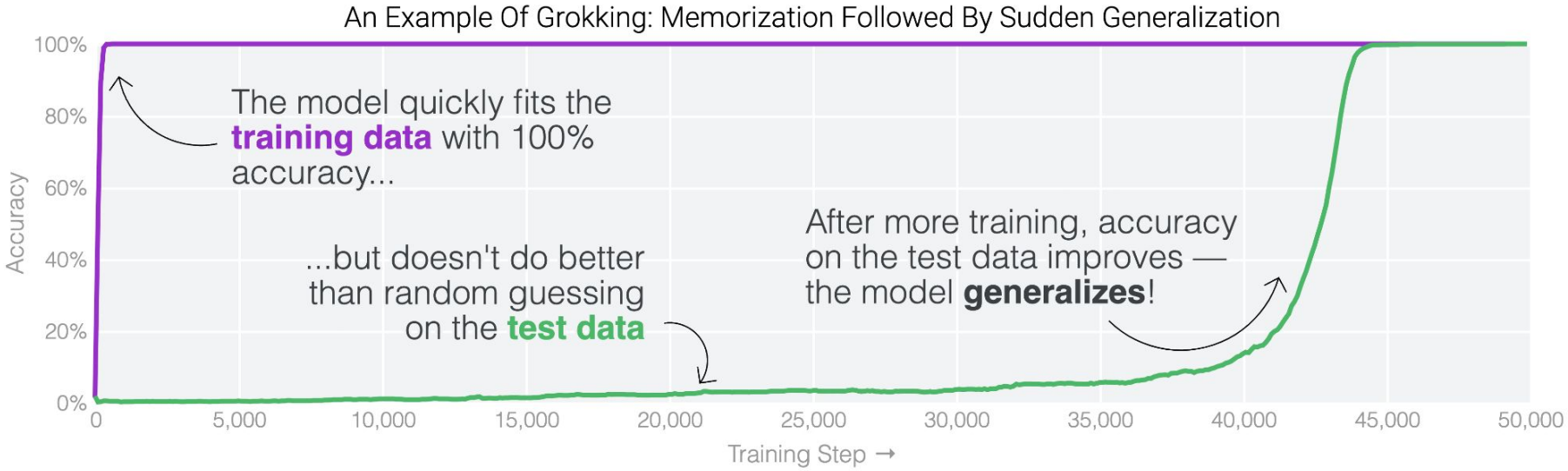


LLM Reasoning

This | breaks | up | words | (| even | | phan | t | a | s | mag | or | ically | long | words |) | into | token | s



LLM Reasoning



https://pair.withgoogle.com/explorables/grokking/

LLM Reasoning

- Pattern matching systems
- Crystallized skills
- No System 2 thinking
- Not capable of reasoning

Forest of Jumbled Thoughts Prompting: An Ultra General Way to use LLMs for Solving Planning, Reasoning, World Peace and Climate Change Tasks

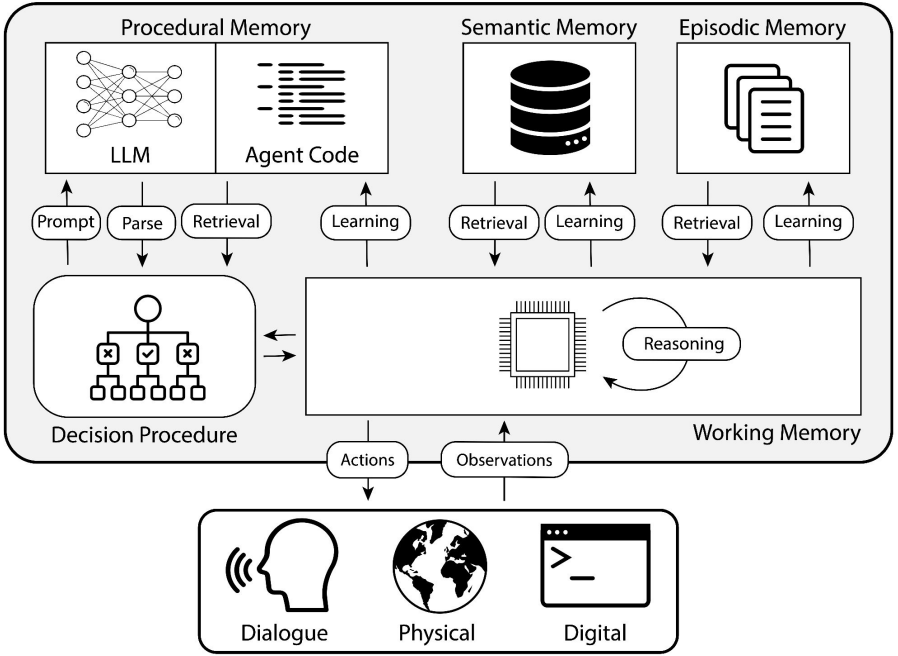
Subbarao Kambhampati
School of Computing & AI
Arizona State University, Tempe.
rao@asu.edu

Abstract

Intrigued by the claims of emergent planning and reasoning capabilities in LLMs, especially in the presence of bright AI graduate students, we have set out to develop the ultimate prompting technique. Our aim is to generalize the chain of thought, circle of thought, tree of thought and graph of thought prompting techniques to a whole another plane. Our "**Forest of Jumbled Thoughts Prompting**" (FJTP) technique is very general, and only requires repeatedly browbeating the LLM to do better by nudging it towards the correct answer. In our experiments on GPT4.5 (that we had got early access to, thanks to our recent investment in OpenAI), we show that our FJTP technique works like a (slow) charm on a variety of planning, reasoning, world peace and climate change tasks. We prove, by reduction to Rube Goldberg Machines, that the FJTP eventually makes LLM "solve" any problem for which the prompting graduate students know the answer. Our proof is general and only assumes an abundant budget for GPT4 API access (or, alternately, co-authors with free access to Palm). The underlying back-to-the-basics "system 2" search that FJTP induces avoids any GOFAI search technology that may need access to things other than LLMs and graduate students. We further show that the solutions that the LLM produces are *exactly the ones the grad students prompt it to produce*–thus ensuring the interpretability and explainability of the solutions generated. We speculate that the awe-inspiring generality of this FJTP prompting technique will eventually make LLMs overcome even their dreaded fear of numbers–and allow them to do arithmetic, thus obviating the need for those *costly* calculators.

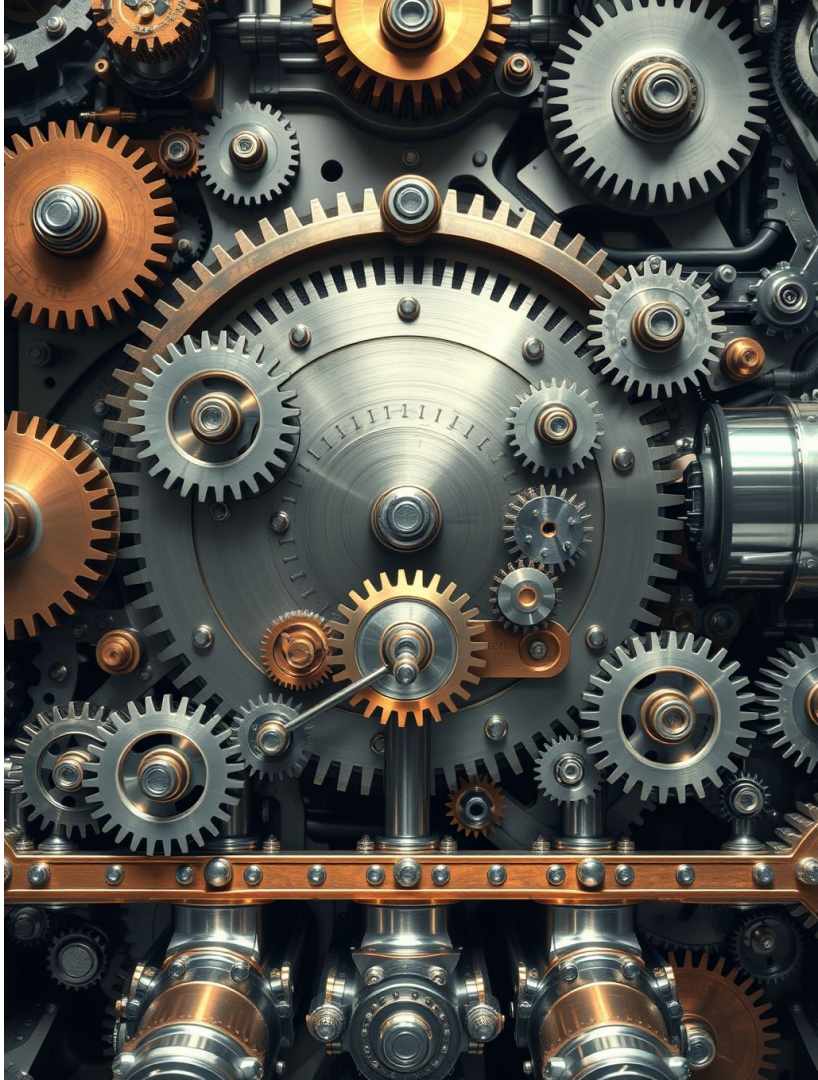
Agentic Reasoning

- Chain-of-Thought
- Cognitive architectures increase reasoning capability
    - LLM Modulo Framework
    - CoALA
    - NEOLAF
- Currently not solved



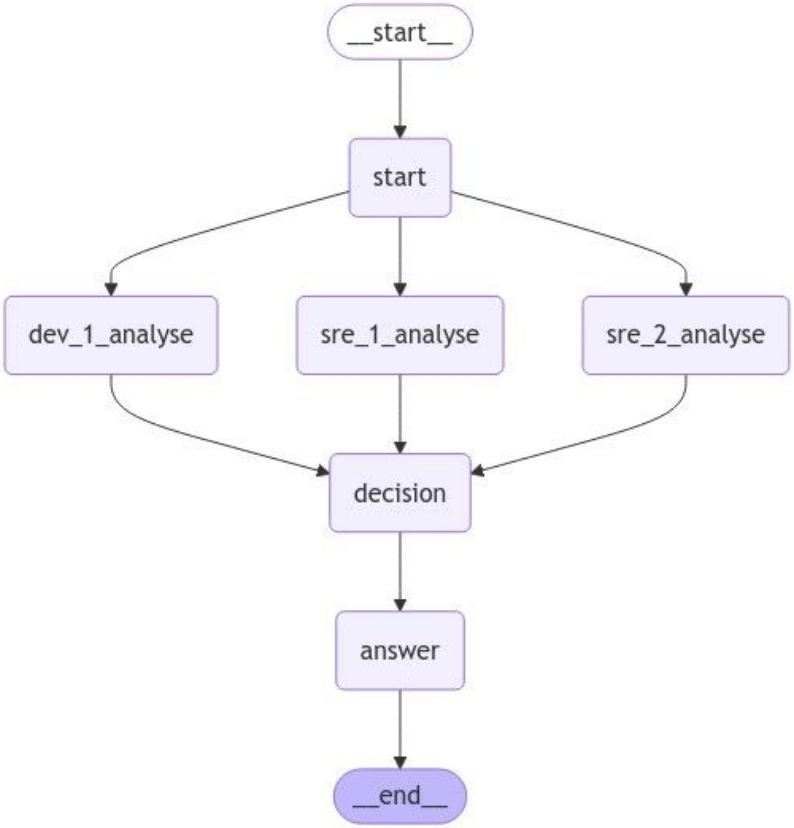
Current Reasoning State

- Domain-specific cognitive model
- Code execution flow models domain specific challenges
- Implemented control points, thought step limits, and plan step limits
- Agent-specific frugal use of tools
- “Reflective-Chain-ReACT”



Multi-Agent Graph

- Multi-agent team emulating an on-call environment
- Task specific agents
- 2 SRE's - CoT and ZeroShot
- 1 Developer - Reflective-Chain-ReACT
- 1 SRE Manager - Agent-as-a-Judge
- 1 Incident Manager - Format prompt



Learnings

- Embrace Non-Determinism
- Pick a single capable LLM
- Pick a single feature, run-book, etc.
- Specialized agent per task or even graph node

Learnings

- **Iterative process**: prompts, hallucination fixes, adding tools, ...
- **Qualitative** evaluation: boolean, integer, class, ...
- Store reasoning traces
- LLMs to evaluate; LLM-as-a-Judge, Agent-as-a-Judge

1. Red Tide
2. Reverts
3. Rapid Iteration
4. Automation
5. Augmentation
6. Future State

Dr. Fix It

- Iterate on agent reasoning
- Implement semantic and episodic memory
- Find commit from stack trace



Dr. Fix It APP 20:53

🔴 **Warning:** around `2024-10-10 10:50:35 UTC` we got `51` of the following error (Kibana):

```
Uncaught Exception: Currencies must be the same (adding EUR to USD)
```

1️⃣ **Possible cause:** 582c85d214f53c3d4f3d404aef3d42c83c1ce213
Deployed `9 minutes` ago at `2024-10-10 10:37:06` in `wpcom-git` by `@[redacted]`

```
Embeds: Bump to 50%

Reviewers: #devops_team!

Differential Revision: https://[redacted]/D163538
```

2️⃣ **Possible cause:** db9b04e03f0e794bd6f222d9d71750ca0a105965
Deployed `4 minutes` ago at `2024-10-10 10:46:50` in `wpcom-git` by `@[redacted]`

```
Removing [redacted] @[redacted] from roundrobin pings

Summary:
For each form submission in agency-engagement-request-form on
https://[redacted].wordpress.com/[redacted]/ - a new
post is created with the submission data as content.
This DIFF removes [redacted] from the pingable people list, so it should only be
pinging two folks from now on.

Test Plan:
- Open https://[redacted].wordpress.com/[redacted]/
- Fill the form and submit
- Confirm that no errors appear and the form submission summary is shown
- CHeck that a new Post was created and that it has all the data submitted in
the form
- Please consider that testing this (even locally) will create a new post and
will tag someone (unless you change the username being tagged in the post)

Differential Revision: https://[redacted]/D163537
```

3️⃣ **Possible cause:** d5f350b11e5ad955f7ed91d7e101e265d664c399
Deployed `4 minutes` ago at `2024-10-10 10:48:15` in `wpcom-git` by `@[redacted]`

```
Revert "Embeds: Bump to 50%"

This reverts commit 582c85d214f53c3d4f3d404aef3d42c83c1ce213.
```

✅ 1

Future

- Kalman filtered agent reasoning distributions?
- Tree-of-swarming-agents with a new search algorithm?
- Stochastic determinism?
- … ?


HEY! GET BACK TO WORK!
Agents finding root cause!
OH. CARRY ON.

AUTOMATTIC

Any questions?