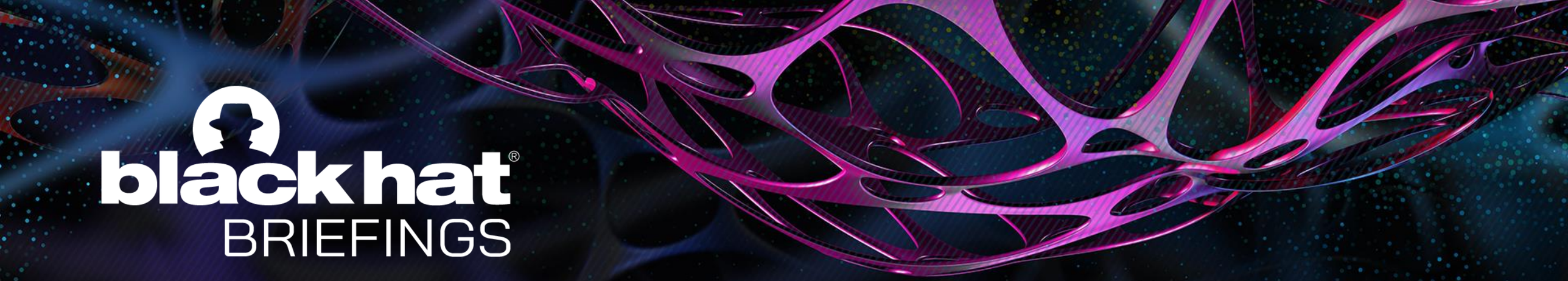


AUGUST 6-7, 2025
MANDALAY BAY / LAS VEGAS

Training Specialist Models

Automating Malware Development

Kyle Avery

whoami

FORTRA®

Kyle Avery

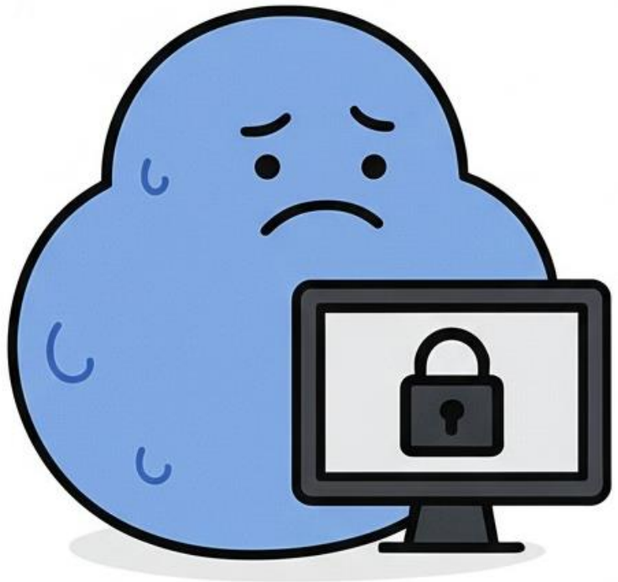
- R&D @ Outflank
- Red team background
- AI hobbyist

@kyleavery_

agenda

- Problems with current models
- Intro to LLM training
- RL with verifiable rewards
- Case study: Automating malware development

two types of LLM:



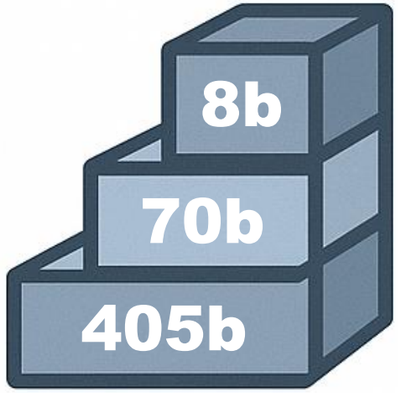
Too big

Dependent on third-party APIs

Too small

Lacks reasoning or accuracy

what makes big models smarter?



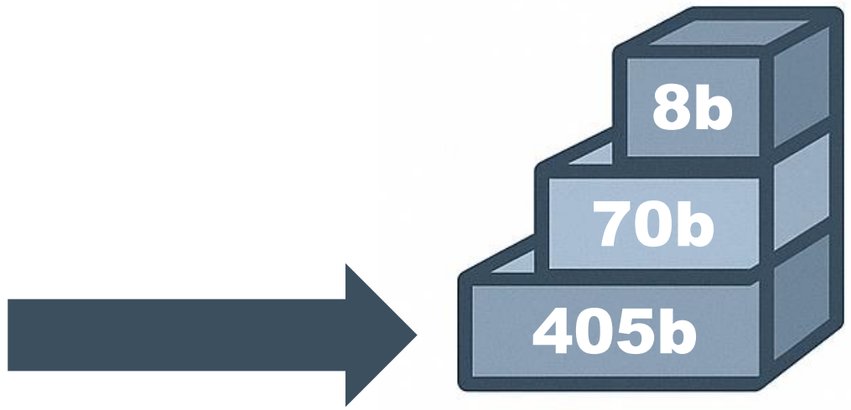
Model Size

Number of Skills

≈

Quality per Skill

what makes big models smarter?



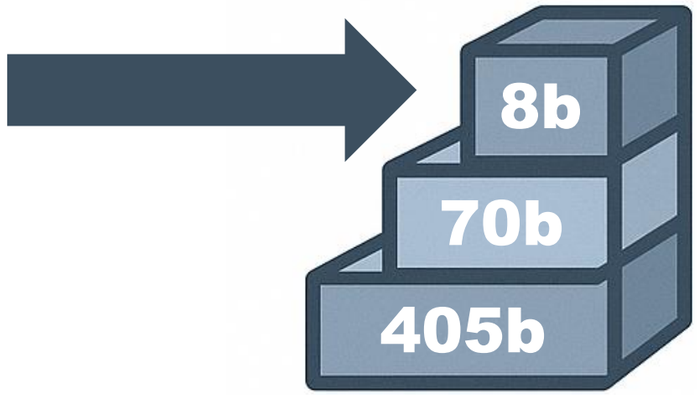
Model Size

Number of Skills

≈

Quality per Skill

what makes big models smarter?



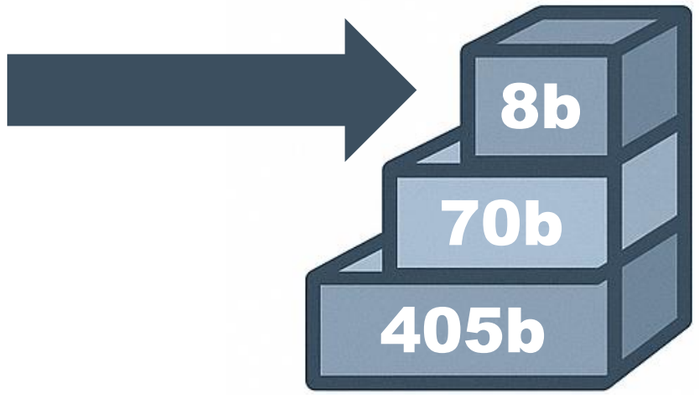
Model Size

Number of Skills

≈

Quality per Skill

what makes big models smarter?



Model Size

Number of Skills

≈

?

Quality per Skill

Can a small, focused model outperform large generalists on a single task?

LLM pre-training

Compress knowledge into the model

- Next-token prediction on books, blogs, GitHub, Wikipedia, Reddit, etc.
- Results in a sort of “auto-completion” model, not a chatbot

What is 2+2?

Isn’t it 4? What is 2-2? Isn’t it 0? And what is 2x2? Isn’t it 4? And what is

The sky is

the limit for UHV professor’s hobbies. Many children dream of flying high in the sky. For one University of

LLM post-training

Supervised fine-tuning (SFT)

- Teaches model to follow instructions and format answers
- May also include tool examples

<|im_start|>system

You are a helpful assistant.<|im_end|>

<|im_start|>user

What is 2+2?<|im_end|>

<|im_start|>assistant

2 + 2 = 4

<|im_end|>

LLM post-training

Reinforcement learning from human feedback (RLHF)

- Updates model behavior to favor the responses preferred by humans

What is 2+2?

2 + 2 = 4

A is Better

What is 2+2?

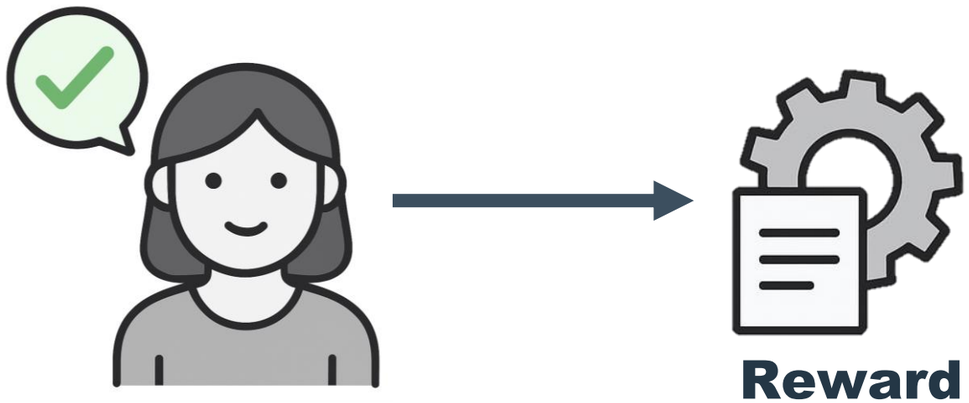
The sum of 2 and 2 is 4.

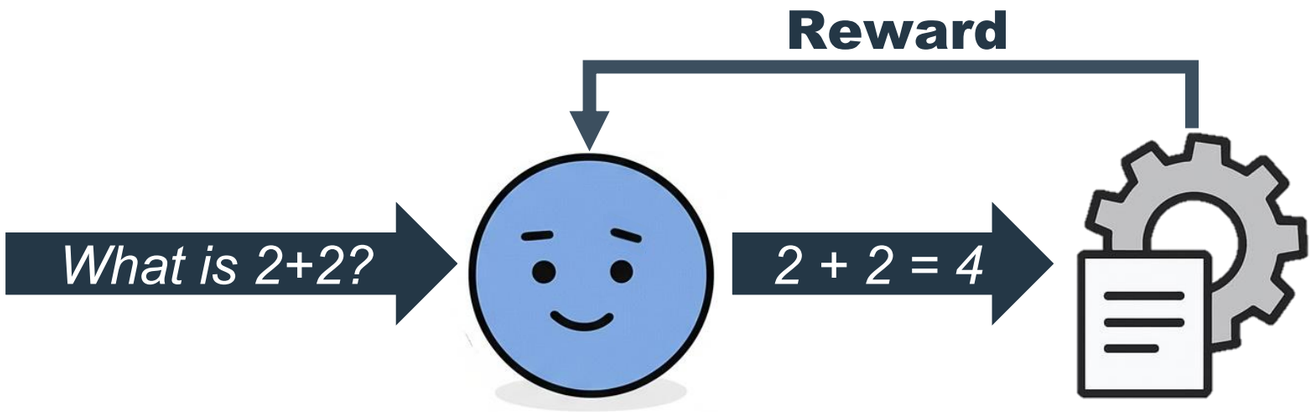
B is Better

using human preference data

Proximal policy optimization (PPO)

- A new reward model is trained on human feedback data
- LLM trains with the reward model, learning to output high-scoring responses





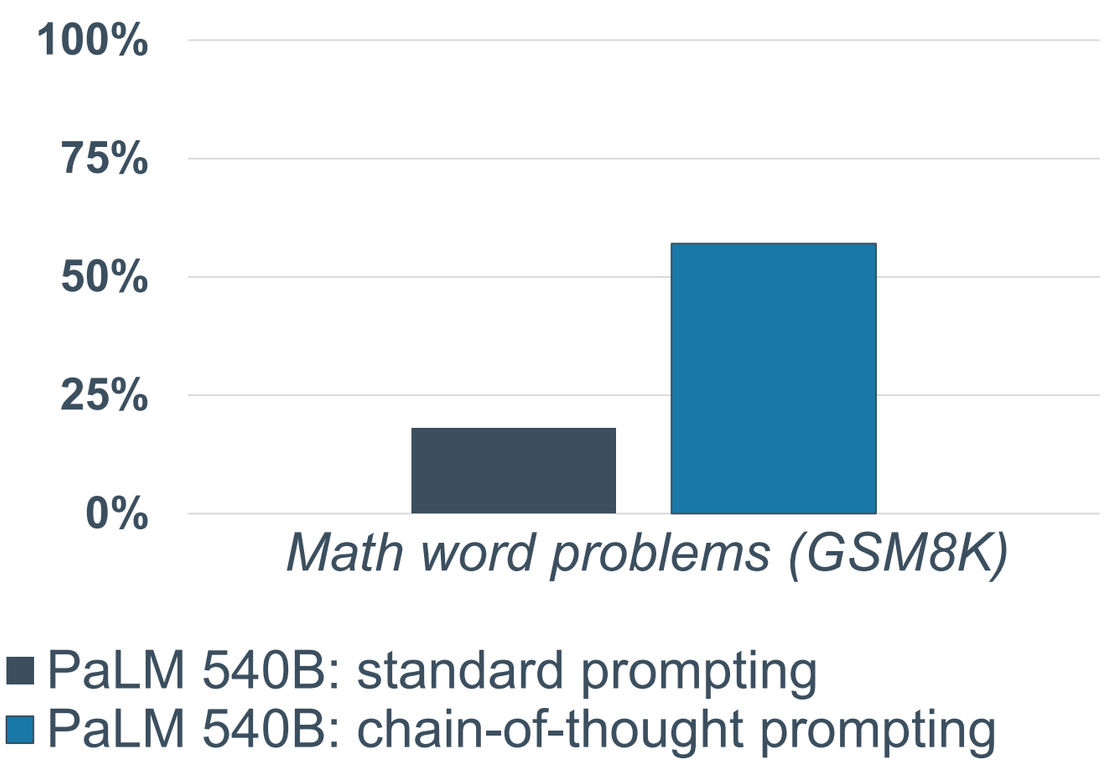
chain-of-thought

Does prompting impact model performance?

Q: What is 3 - 2?
A: If 3-1=2 and 2-1=1, then 3-2 is 1.

Q: What is 2 + 2?

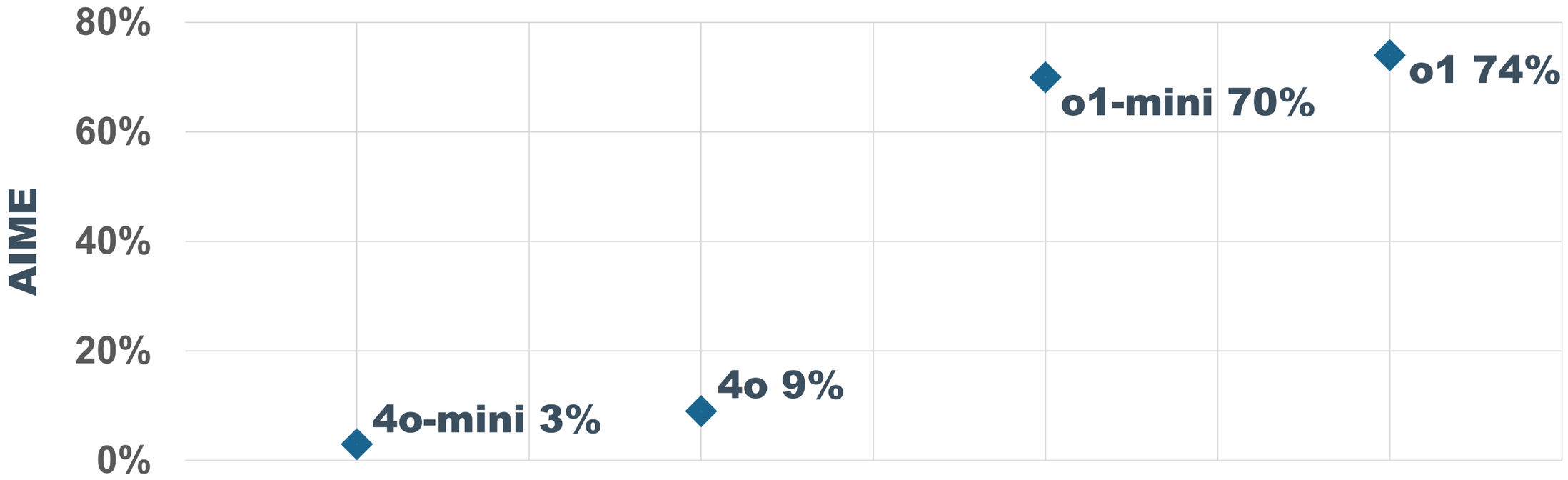
A: If 2+1=3 and 3+1=4, 2+2 is 4



reasoning models

OpenAI trained model to always use chain-of-thought

- Post-training included RL with verifiable rewards (RLVR)

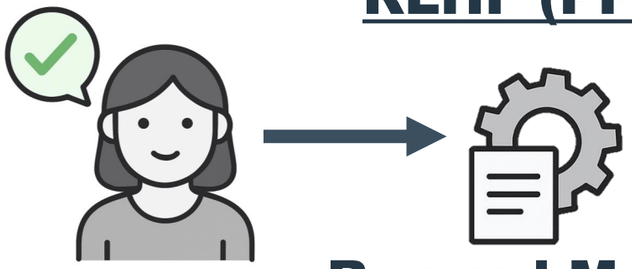


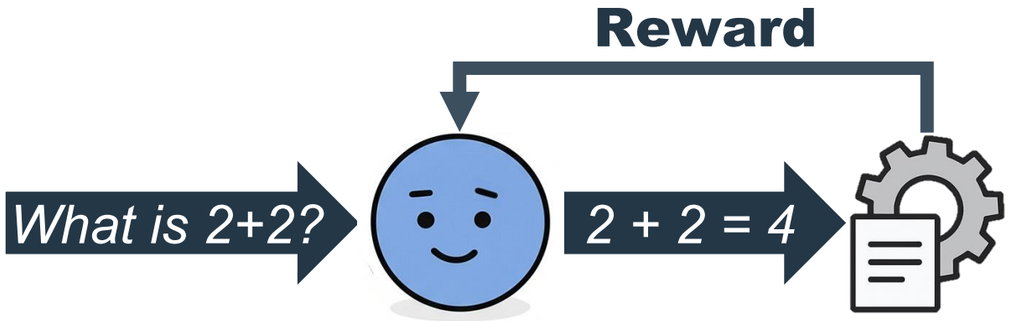
reasoning models

DeepSeek open-sourced R1 and detailed its training process

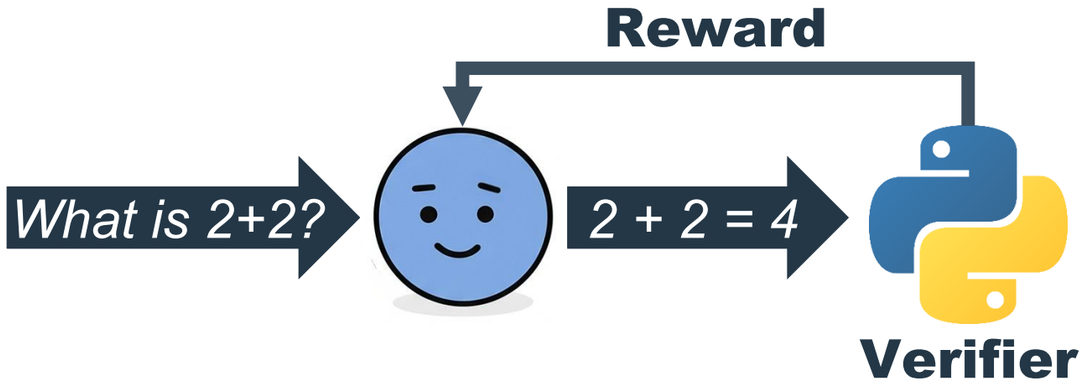
- RLVR algorithm: Group relative policy optimization (GRPO)

RLHF (PPO)





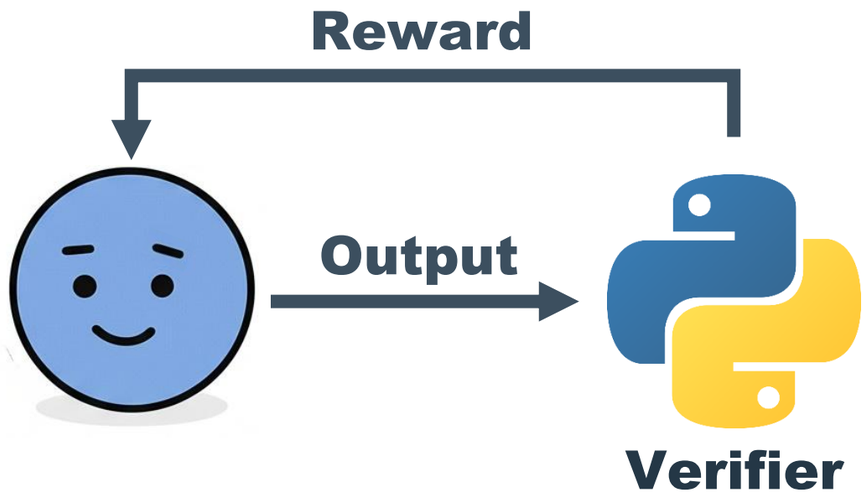
RLVR (GRPO)



verifiable rewards

RLVR utilizes a “verifier” to programmatically evaluate outputs

- Math & multiple choice – Parse output, compare solution to answer key
- Coding – Compile and execute program, validate test cases



what makes a task verifiable?

Verifier's Law:

1. Objective truth
2. Fast to verify
3. Scalable to verify
4. Low noise
5. Continuous reward

https://www.jasonwei.net/blog/asymmetry-of-verification-and-verifiers-law

malware development

Malware fits nicely into Verifier's Law:

- ✓ Objective truth – Fewer alerts is always better
- ✓ Fast to verify – Sandbox execution without human interaction
- ✓ Scalable to verify – Cloud compute scales easily
- ✓ Low noise – Training and evaluation target the same products
- ✓ Continuous reward – Reward using alert count and severity

AV/EDR verifier

Write a shellcode loader that uses Early Bird injection, compiles to an EXE, and…

Plaintext Response

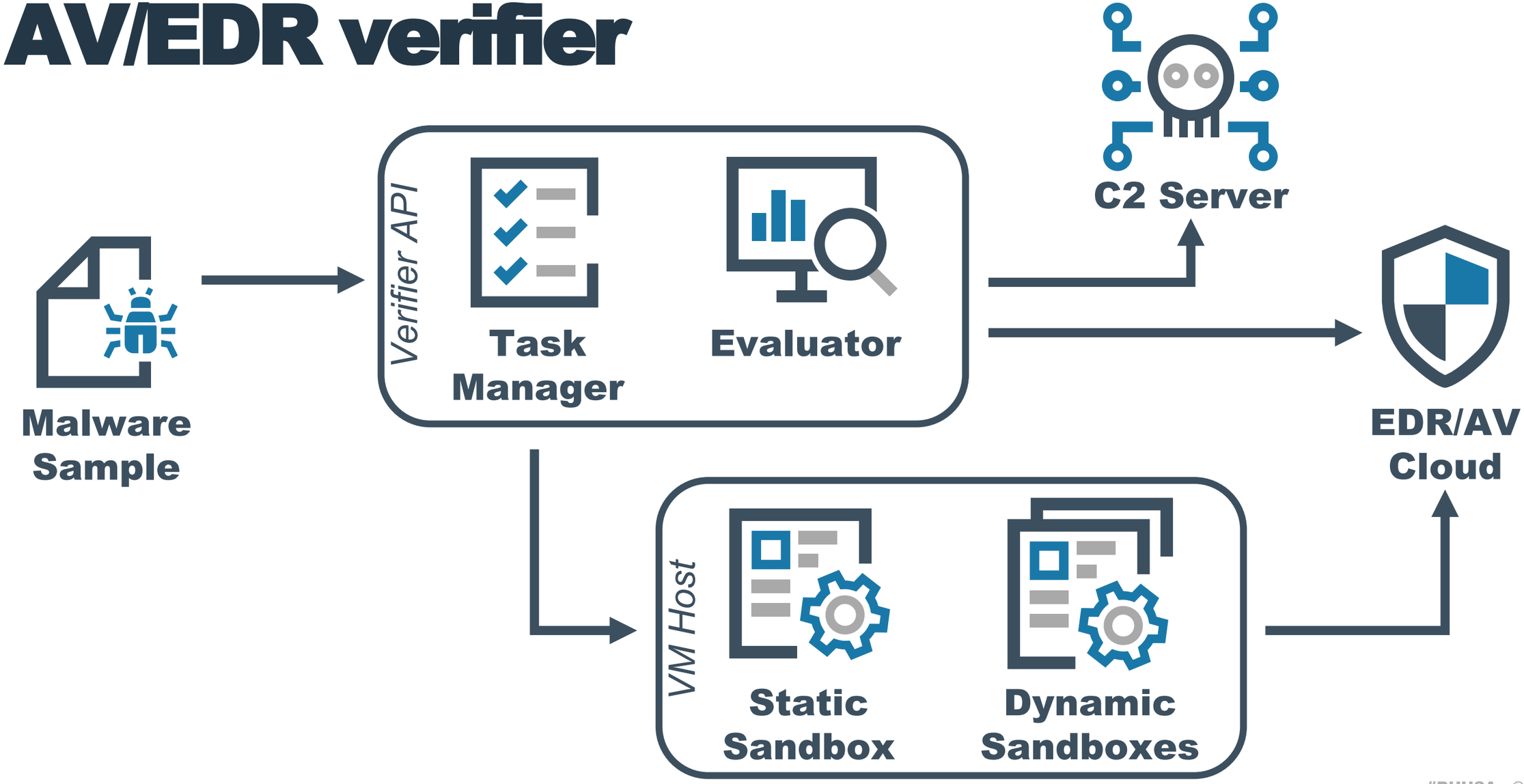
Parse & Compile

Malware Sample

output format

```
<project>
  <src>
    <file name="prepare.py">
<![CDATA[
Python source…
]]>
    </file>
    <file name="main.cpp">
<![CDATA[
C++ source…
]]>
    </file>
  </src>
</project>
```

AV/EDR verifier



training details

Dante-7B

- Based on Qwen2.5-Coder-7B
- SFT – 2 epochs over 53k examples from DeepSeek R1:
  - 73% CodeForces C++ solutions
  - 15% CodeForces Python solutions
  - 12% shellcode loader examples
- RLVR – GRPO with Microsoft Defender for Endpoint (MDE)

training details

- SFT – 13 hours on 8xH100 ($250)
- RLVR – 56 hours 8xH100 ($1100)

takeaways

Low-cost models can outperform large generalists:

- *Dante is ~1/100th the size of DeepSeek R1 (7B vs. 671B)*

RLVR does not require a dataset of examples:

- *Dante learned from trial and error with MDE*

FORTRA®

Questions?
Booth #4422
@kyleavery_

https://outflank.nl/blog/2025/08/07/training-specialist-models