Vulnerability Research in the Age of AI

keynote @ VXCON 2024

Alisa Esage, Zero Day Engineering Project

Alisa Esage

Founder & CEO

Zero Day Engineering Research & Training

@zerodaytraining

Independent Hacker

Advanced Reverse Engineering & Exploit Development @alisaesage

Cybersecurity Veteran

Since 1999

“I am an engineer, Jim. When I see things, I see them.”

Dagny Taggart

The Plan

deep insight to see through hype cycles and technology limitations

Part I. Core of AI technology

Part II. Unique challenges of VX

Part III. Analysis

Conclusions

Part I. Core of AI technology

AI vs. Cybersecurity

background

- AI has a long history in Cybersecurity
    - Attacking AI models
    - Applications to core Cybersecurity challenges
    - Defense & Offense
- Emergence of genai & LLM changes everything
    - Pattern recognition becomes advanced enough
    - “Emergence”?

AI Evolution

ML >> DL >> LLM

- Early research: 1940s-1960s
  - Artificial neuron, Turing test…
- **Machine Learning**: 1990s-2010s
  - Reinforcement Learning, Deep Blue…
- **Deep learning**: 2010s-2020s
  - Convolutional models & Generative Adversarial Networks
- **Large Language Models**: 2020s-…
  - Transformer model architecture & GPT

Large Language Models

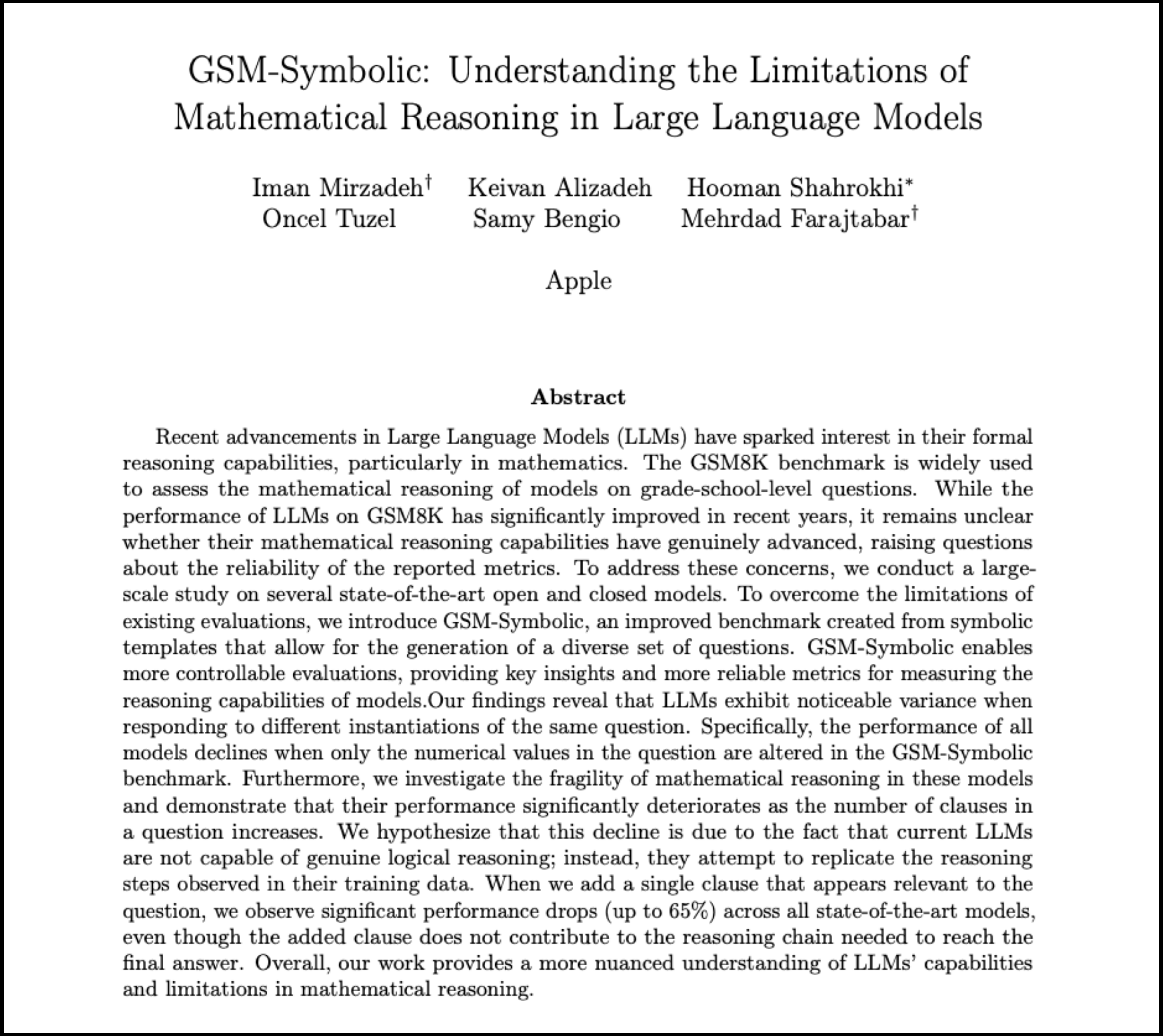
natively optimized for processing language data

- Recognition of highly structured patterns
  - “Attention”, huge training datasets, expensive training
- Generation of highly structured data
  - Based on previously learned patterns
- Ultimately, LLM is a statistical model that works in both directions
  - “Sufficiently advanced technology is indistinguishable from magic”

But can LLMs reason?

not really

- Advanced enough pattern recognition comes off as pseudo “reasoning”
- Data contamination corroborates the confusion
- Smart ways to test this empirically
- How do you know that human level intelligence isn’t just an (even more) sophisticated pattern recognition?



GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models

Iman Mirzadeh[†] Keivan Alizadeh Hooman Shahrokhi[*]
Oncel Tuzel Samy Bengio Mehrdad Farajtabar[†]

Apple

Abstract

Recent advancements in Large Language Models (LLMs) have sparked interest in their formal reasoning capabilities, particularly in mathematics. The GSM8K benchmark is widely used to assess the mathematical reasoning of models on grade-school-level questions. While the performance of LLMs on GSM8K has significantly improved in recent years, it remains unclear whether their mathematical reasoning capabilities have genuinely advanced, raising questions about the reliability of the reported metrics. To address these concerns, we conduct a large-scale study on several state-of-the-art open and closed models. To overcome the limitations of existing evaluations, we introduce GSM-Symbolic, an improved benchmark created from symbolic templates that allow for the generation of a diverse set of questions. GSM-Symbolic enables more controllable evaluations, providing key insights and more reliable metrics for measuring the reasoning capabilities of models.Our findings reveal that LLMs exhibit noticeable variance when responding to different instantiations of the same question. Specifically, the performance of all models declines when only the numerical values in the question are altered in the GSM-Symbolic benchmark. Furthermore, we investigate the fragility of mathematical reasoning in these models and demonstrate that their performance significantly deteriorates as the number of clauses in a question increases. We hypothesize that this decline is due to the fact that current LLMs are not capable of genuine logical reasoning; instead, they attempt to replicate the reasoning steps observed in their training data. When we add a single clause that appears relevant to the question, we observe significant performance drops (up to 65%) across all state-of-the-art models, even though the added clause does not contribute to the reasoning chain needed to reach the final answer. Overall, our work provides a more nuanced understanding of LLMs’ capabilities and limitations in mathematical reasoning.

The big problem

research doesn’t understand how AI works

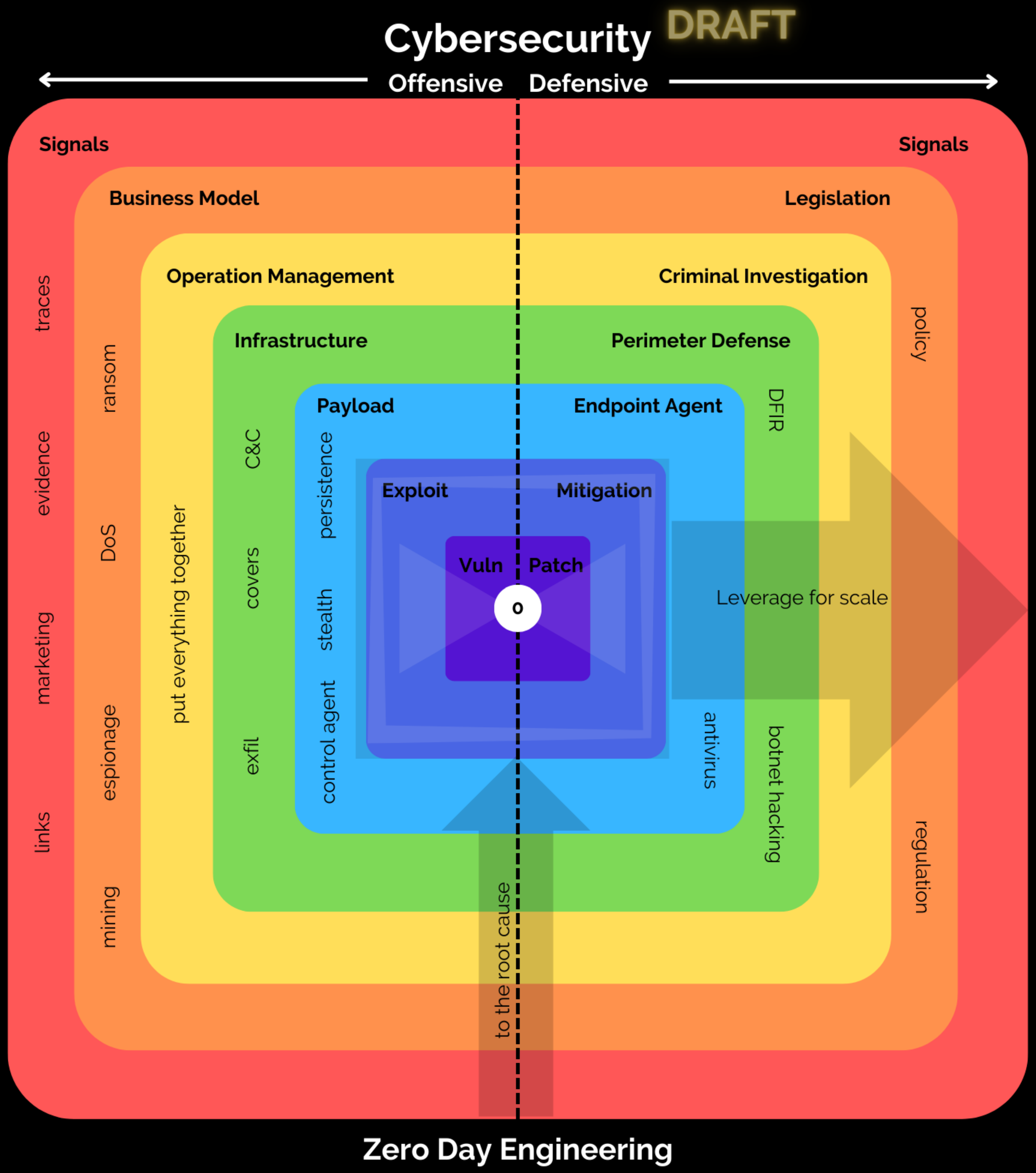
- Dedicated research areas: Interpretability & Model Debugging
- The more advanced an AI model, the less insight is available into its inner workings
- Security risks is one issue
- Lack of grasp on model limits is another
- Rapid advancement driven by empirical feedback loops and hype, how can it be a great way to build?

Part II. Unique challenges of VX

VX vs. Cybersecurity

scope of this talk

- Vulnerability lies at the core of the industry
- It is the foundation on which everything else stands
- Take out vulnerability, and there is no more need for cybersecurity
- Offensive vulnerability research partially solves vulnerabilities
- No full solution is possible



Offensive Vulnerability Research

vulnerability discovery + exploit development

- Vulnerability Research
  - Focus on discovering and understanding vulnerabilities
  - Key driver: **pattern recognition**
  - Prerequisites: knowledge of vulnerability patterns, capability to test hypothesis
  - Scope: unsolved parts

Offensive Vulnerability Research

vulnerability discovery + exploit development

- Exploit Development
  - Focus on manipulating a vulnerability to achieve a desired program state
  - Key driver: **dynamic reasoning** in a tight empirical feedback loop
  - Prerequisites: goal fixation, iterative workflow, **abstract thinking**
  - Scope: non-trivial exploit development

Part III. Analysis

Vulnerability Research

vs. LLM

- Vulnerability Research
  - Focus on discovering and understanding vulnerabilities
  - Key driver: **pattern recognition**
  - Prerequisites: knowledge of vulnerability patterns, capability to test hypothesis
  - Scope: unsolved parts

Exploit Development

vs. LLM

- Exploit Development
  - Focus on manipulating a vulnerability to achieve a desired program state
  - Key driver: **dynamic reasoning** in a tight empirical feedback loop
  - Prerequisites: goal fixation, iterative workflow, **abstract thinking**, specialized practical skills
  - Scope: non-trivial exploit development

Professional relevance

security researchers in post AI world

- Focus on frontiers of security research
  - Non-trivial exploit development
  - Novel vulnerabilities for which the abstract pattern isn’t known
- Focus on AI technology
  - Reverse engineering and introspection is a fundamental problem in AI research
  - Offensive research will eventually drive AI Safety just as it now drives Cybersecurity

Transparency Apocalypse

ultra long term projection

- LLMs are now used at scale to generate code
- They are starting to get used to find bugs in the code
- This will eventually be automated in a loop
- Vulnerabilities go deeper and subtler
- Humans won’t always be able to follow
- Ultra narrow 0-day research margin with superhuman requirements (maybe)
- This projection will manifest unless somebody stops it

“A problem cannot be solved with the same consciousness which created it”

Albert Einstein

Conclusions

- Understand the core of both the proposed tool and the subject
    - See through present limitations of AI technology to make good investment
    - Specialized knowledge of VR+XD is mandatory to build a working solution
- LLMs are natively great for Vulnerability Discovery
    - Problems with it are fundamental AI research problems
- LLMs are natively not so great for non-trivial Exploit Development
    - Abstract reasoning capability requires another leap in AI evolution

Hacking is everything that AI is not

It’s the one job that AI will actualize and enrich rather than marginalize and eliminate

If you do it right