



Operationalizing AI:

The Race for an **Unfair Business Advantage**

Damon Edwards
@damonedwards

We put the **AI** in **operations**.

AIOps | Process Automation | Incident Response | Customer Service Ops

PagerDuty
Operations
Cloud

OUTFRONT

158192

6896

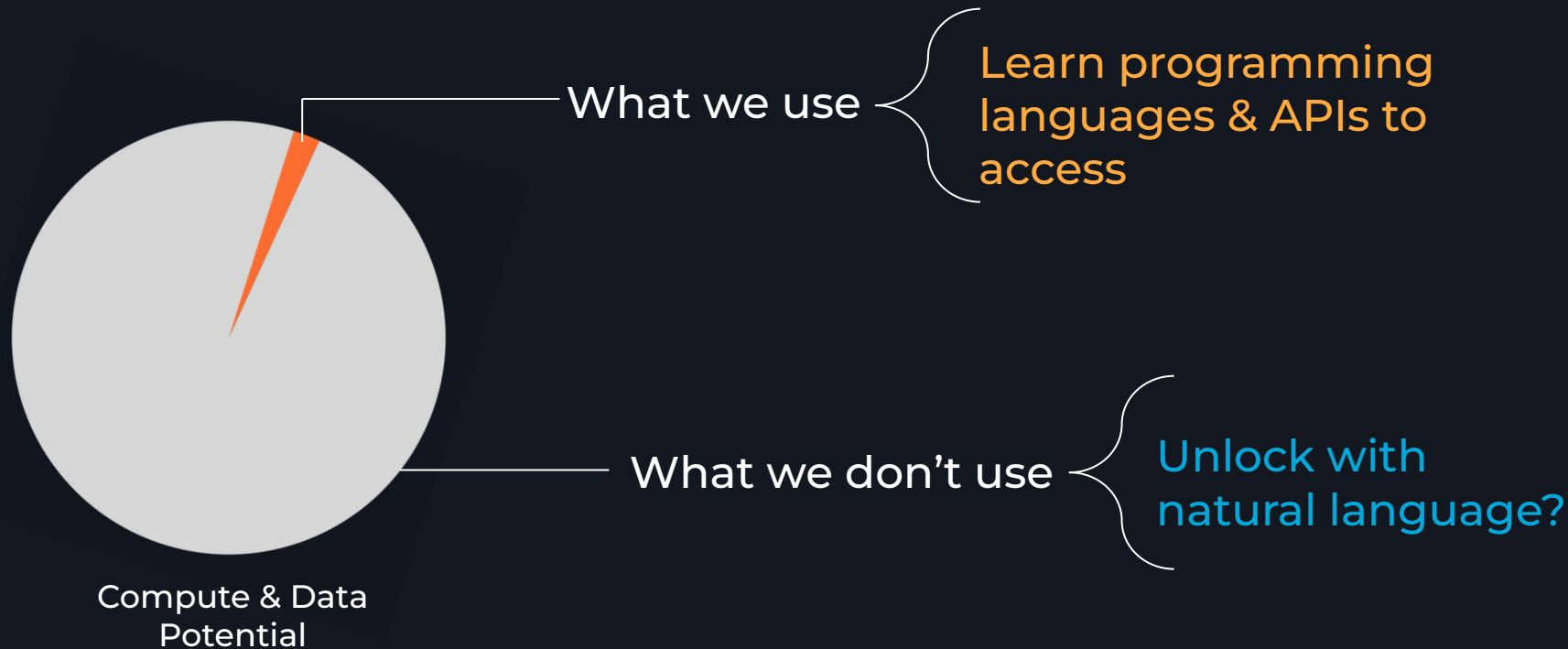
I asked colleagues around the industry...

Q: Why such a big bet on AI?

A: Because we can't afford to be wrong



Compute & Data
Potential



What can we do with that? (external product focus)

- Revisit ideas that were previously thought to be too hard or too costly
 - Inputs or user intentions had too much variability / unpredictability
 - Underlying data had too much variability or lack of structure
- Solving at a fraction of the cost previously expected, gains **structural advantage** over competitors who can't

What can we do with that? (internal focus)

- **Margins** are life!
- Change the game for you biggest cost... *your people*
- Repetitive toil, communication ceremonies, etc.
- Previously too hard to automate... missing linguistic layer
- Today: Tasks... 30% improvement?
- Tomorrow: Roles... 300%? 3000%?
- Large margins, relative to peers, is a **structural advantage**

Large enough structural advantages
turn into an *unfair business advantage*.

Possibility of of being on either side of an
unfair business advantage?

You have to move... *Now*.

But, seriously, why the urgency?

- Web →→→ Cost additive for future strategic benefit
- Mobile →→→ Cost additive for future strategic benefit
- AI →→→ Cost *deflationary* for *immediate* strategic benefit

When one in a sector figures it out, everyone else must as well.

Someone is going to lead your sector, *why not your company?*

Imperative: Get your company moving!

OK, but how?

(some early lessons and observations)

Jury is still out on “AI Committees”

- Haven't seen it work yet
- It's a committee! (you know how that goes)
- Who do you put on the committee? (old experts \neq new experts)
- Weak execution ownership

Getting a company into discovery mode

- Discovery of the unknown comes from the people closest to the work
- Discovery & learning has to be baked into everyone's work
- Challenges the assumption that “someone else will figure it out for us”

Getting a company into discovery mode

Option A: Anoint specific teams

Pros

- Pick (hopefully) easy wins
- Focus/swarm resources
- Keep the rest of the org on plan

Cons

- Only single/few teams moving
- Maybe the new stars / best uses are somewhere else in org
- Osmosis is really slow
- Telegraphs that “figuring it out” is someone else’s problem

Getting a company into discovery mode

Option B: Company-wide prototype sprint (aka hackathon)

Pros

- *Everybody* gets activated
- New ideas & new stars will emerge from within
- Sends a bold message

Cons

- Finance and ops (in the COO sense) will freak out at the idea
- Need to overcome people's fear of getting their "hands dirty"

Getting a company into discovery mode

If you are going with Option B...

- Don't let it devolve into a free for all of tangents
- Pick a stack for everyone / Provide patterns
- Encourage people to learn about the stack in advance

Example of
a "stack"

Prompt Engineering
Python
Notebooks (e.g Google Colaboratory)
App Patterns (e.g. LangChain)
Vector DB (e.g. Chroma + one with persistence)
LLM (e.g. OpenAI)

*Beware the trap of
the vendor paradox!*

← Ok for some to just be engaging
with prompt engineering

Focus where the company actually need it

You are operationalizing AI... not building AI

Prompt Engineering
+ chaining APIs

→ Prove it here!

Fine-tune a model

→ Be very skeptical

~~Train a Model~~

→ Stop the meeting

Surprise consensus: It's easier than expected!

Perplexity.ai – Proprietary Service

critical analysis of Google "We Have No Moat, And Neither Does OpenAI"
Leaked Internal Google Document Claims Open Source AI Will Outcompete Google and OpenAI

* Cópilot

Q Searching web

critical analysis Google "We Have No Moat, And Neither Does OpenAI" leaked internal document

OpenAI's outsource Google leaked internal document analysis

Considering 15 sources

📄 Sources

- Google "We Have No Moat, And Neither Does..."
semianalysts - 1
- Leaked Google document reveals open source AI...
emergentmind - 2
- Google "We have no moat, and neither does..."
linkedin - 3
- Leaked Internal Google Document Claims Open...
hackaday - 4
- Google "We have no moat, and neither does OpenAI"...
- ID Google "We Have No Moat, And Neither Does..."
reddit - 6
- Leaked Google Memo: "We Have No Moat, and..."
marketinsights - 7
- View 6 more

📄 Answer

A leaked internal Google document claims that open-source AI models are advancing rapidly and may outpace both Google and OpenAI. The document suggests that Google should embrace open-source AI, prioritize third-party integrations, and reconsider the focus on large models. The anonymous Google researcher who leaked the document believes that neither Google nor OpenAI is positioned to win the AI "arms race" due to the rapid advancements made by open-source AI projects.

One of the key factors contributing to the success of open-source AI projects is Low-Rank Adaptation (LoRA), a technique that allows model fine-tuning at a fraction of the cost and time by reducing the size of update matrices. The document argues that Google and OpenAI do not have a "secret sauce" that makes their approaches better than what the wider community can come up with. The researcher suggests that Google should establish itself as a leader in the open-source community, cooperate with the broader conversation, and potentially publish model weights for small ULM variants.

The implications of this leaked document could significantly impact Google and OpenAI's future business strategies, as both companies may need to reevaluate their approach to AI research and development in light of the rapid advancements made by open-source AI projects.

Learned how to re-create in 1.5 days (with python examples from Github, LangChain, GPT API)

Critical Analysis of Leaked Google Document: "We Have No Moat, And Neither Does OpenAI"

The contemporary field of artificial intelligence (AI) has been stirred by the recent leak of an internal Google document penned by a Google researcher. The memo, while not an official statement from the company, suggests a drastic shift in the AI landscape - the imminent threat from open-source AI. The document states that Google and its AI industry competitor, OpenAI, have no "moat" in the AI technology domain, referring to the lack of a unique and constant competitive edge. This report provides a rigorous examination of the implications raised by the leaked document, the responses thus far, and the broader landscape of the AI industry.

The Clarion Call: No Moat in AI

The leaked memo alludes to the rapid advancements of open-source AI projects, outpacing proprietary models developed by tech giants like Google and OpenAI. The document also insinuates that the "quality gap" between open and proprietary models is dwindling. Factors such as speed, customization, privacy, and capability have been spotlighted as areas where open-source projects are excelling. Tangible exemplifications of this have been noted with models like Meta's LLaMA language model and Vicuna-13B, an open-source chatbot. Both have evidenced the potential of open-source AI to challenge more established players. However, the text provides no quantifiable data to support these assertions, leaving the actual degree of this impact open to interpretation.

Response from Google and OpenAI

Contrarily, Demis Hassabis, CEO of Google's DeepMind, authenticated the leaked memo but disagreed with its conclusions. Hassabis appeared optimistic about Google's stand in the AI industry, pointing to the company's innovative researchers and the potential for progress through collaborations between Google Brain and Google DeepMind teams. No specific counter-arguments, numbers, or additional expert opinions were provided to refute the contents of the memo. Also, there has been no formal response from OpenAI, thereby adding a level of ambiguity to the situation.

Open-source AI: Boon and Bane

Interestingly, open-source AI is represented as both an opportunity and a challenge. The mounting momentum of open-source AI projects, with breakthroughs like the Low Rank Adaptation technique, and reduced entry barriers have emphasized the democratizing potential of open-source AI. Yet, there are also concerns about potential EU regulations on open-source models, which could impede the release of open-source AI in Europe. Here, the text lacks specifics on what these regulations could entail and how they might shift the AI dynamic.

On the flip side, while the open-source AI revolution poses potential threats to AI stalwarts like Google and OpenAI, it simultaneously prompts the necessity for these organizations to reassess their competitive strategies. The documents

stress the need to embrace collaborative measures, learning from the innovations in the thriving open-source community, and staying open to external collaborations to retain their relevance in the AI market.

Beyond Openness: A Matter of Dominance

The Wired.com article brings a broader perspective to this discourse by highlighting the challenges of true openness in AI. The researchers argue that ownership of software frameworks, the high cost of computing power, secrecy around training data, and the lack of available expertise to fine-tune models concentrate power among large corporations. In this context, the open AI model may consolidate authority rather than lead to democratization. Memorably, it proposes the need for regulatory measures to avoid technological dominance by large corporations, especially in sensitive domains like healthcare, finance, education, and the workplace. Here again, further context or specifics could provide a more comprehensive understanding of these challenges.

Conclusion

The document leak may be seen as a hypothesis on the future trajectory of the AI race, highlighting that existing AI entities might not be as infallible as presumed. However, the underlying analysis remains somewhat speculative in the absence of concrete statistical data to support these assertions. Therefore, while it provokes thoughtful discourse on the future of AI, the suggestions of this document should be considered as one of the many perspectives in the dynamically evolving AI landscape.

References

- SemiAnalysis - Google "We Have No Moat, And Neither Does OpenAI"
- Artisan - Leaked Google Memo Claiming "We Have No Moat, And Neither Does OpenAI" Shakes
- LinkedIn - Google "We Have No Moat, And Neither Does Open AI"
- The Guardian - Google Engineer: Open Source Technology AI OpenAI
- UX Design Bootcamp - Seizing The Open Source AI Opportunity
- CMSWire - The Open Source Revolution Challenging AI Giants
- The Verge - Google Memo Moat AI Leak
- DataCamp - OpenAI Google AI Data Science
- DeepChecks - OpenAI's ChatGPT vs. Google's Bard AI: A Comparative Analysis
- Wired - The Myth of Open Source AI

Experiment by Joseph Enochs (evtcorp.com)

@damonedwards

What to expect

Table stakes today - do it now

- Coding co-pilots
- Summarization
- Synthesize, manipulate, create new output
- Generate code
- “Chat with documents/data” (RAG Chatbots)

Still emerging

- Multimodal

Not ready for prime time

- Agents (the truly open-ended, goal-oriented type)

Perform
Tasks
(\$)

Perform
Roles
(\$\$\$\$)

Don't put unreal expectations on Data Science

- In most companies, Data Science team didn't ship or operate products... don't suddenly expect them to
- DevOps is as foreign to them as AI is to us
- They are way out numbered and everyone wants their time
- A promising marriage: Data Science ↔ Platform Engineering
 - AI experts plus delivery & scale expertise
 - Deliver the “dial-tone” that other teams can leverage at scale

Shadow AI is inevitable

The Rise of Shadow AI

Published On: July 5, 2023 By John Willis



"Before AI is smarter than us, I think the people developing it should be encouraged to put a lot of work into understanding how it might try and take control away," – Geoffrey Hinton, Godfather of A.I.

- The rise of Generative AI and ChatGPT is leading to a new version of Shadow IT with increased security and compliance risks.
- Generative AI large language models are complex and opaque, with few experts understanding how they work, leading to increased security risks.
- Organizations must learn to manage the increased risks of using open source software in generative AI solutions.
- IT leaders need to implement standardized internal training programs, standardize the use of LLMs and develop secure LLM delivery workflows.
- New policies and observability patterns must be developed to consider vulnerability scans of LLM code and dependencies and detect anomalies, breaches and data leaks.
- Recent innovations like retrieval augmentation tools have reduced LLM hallucinations and allowed for safer

John Willis - <https://techstrong.ai/aiops/the-rise-of-shadow-ai/>

It's going to happen

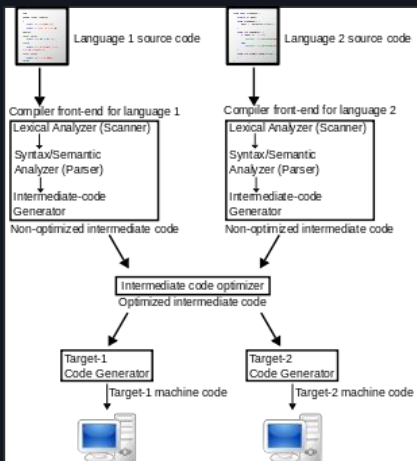
- People avoiding bottlenecks
- Enthusiasm + Need

How do we *safely* harness it?

Who will be your enthusiasts? detractors?

- Everyone surprised at where specific people's opinions fall
- I have a partial theory:

Dev POV: "Non-determinism is a bug!"



Ops POV: "Non-determinism is reality!"

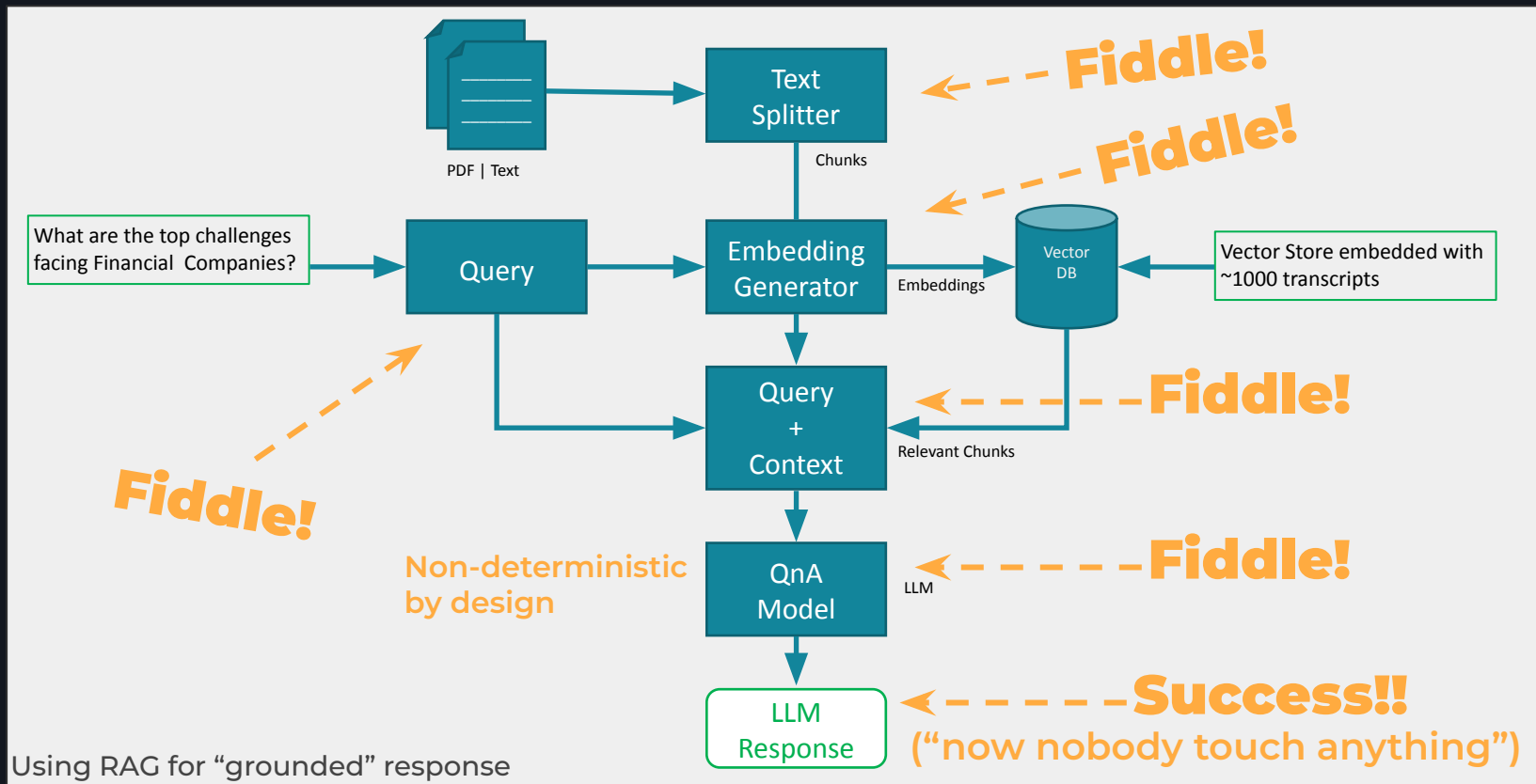


@SoberBuildEng

<https://jpaulreed.com/thoughts/dev-ops-and-determinism.html>

@damonedwards

Generative AI is a lot of fiddling!



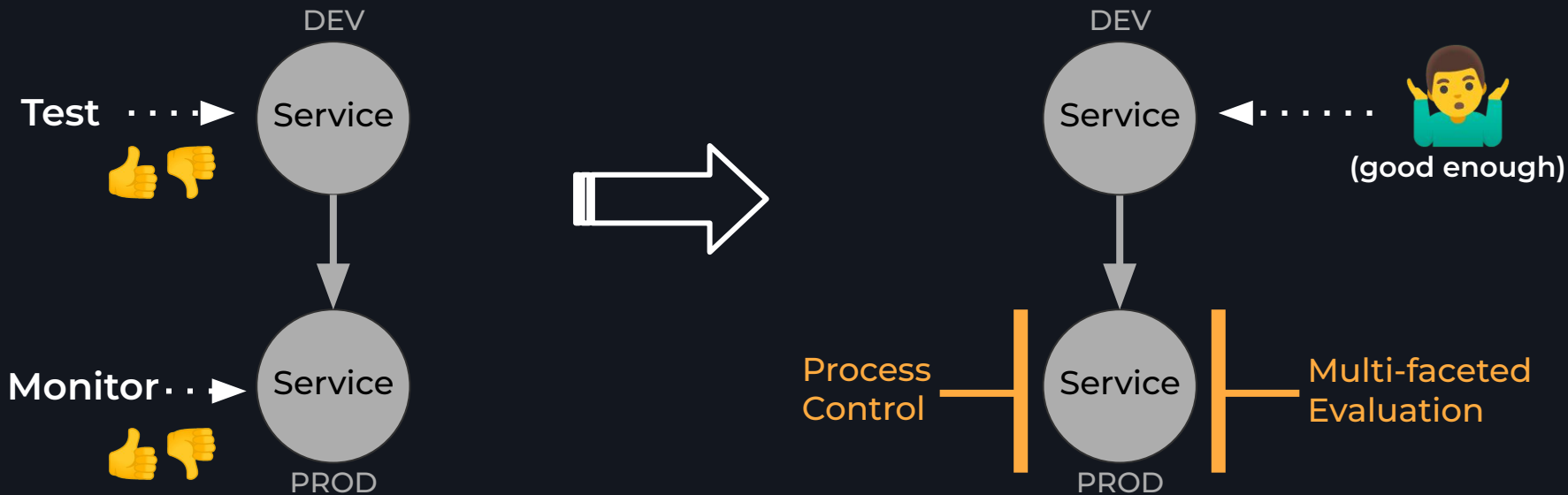
Using RAG for "grounded" response

Diagram Credit: GenAI Hackathon by TechStrong.AI

@damonedwards

“Quality Assurance” looks a lot Different

- 70+ years of trying to root out non-determinism
- Rethink for when *non-deterministic by design*



Changing how we think about UX

Tasks & Data Views



Conversations?



Beware of prematurely optimizing for cost

- *Many tokens used during conversations*
- *Many tokens used to control and evaluate outputs*
- *Prematurely worrying about token usage will **stunt your organization's discovery and adoption***
- *Trends showing hyper-competition around foundation models
– substantial price decreases seem inevitable*

How I can use your help... Let's Talk!

Working on a new **community effort** to document and share patterns and lessons learned around operationalizing AI

damon@operationalizing.ai