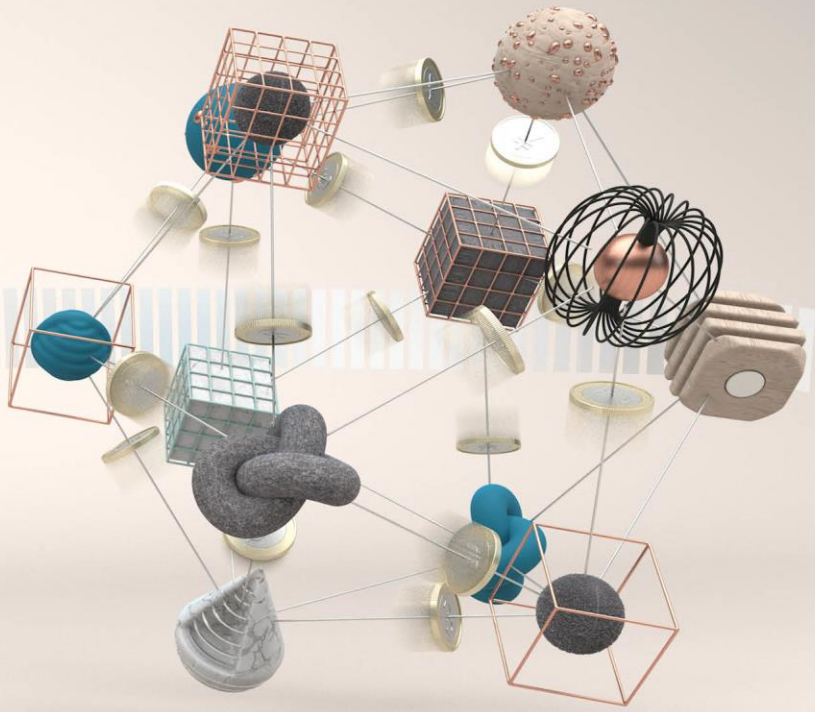
BANKING CIRCLE

What OR can learn from putting LLMs into production



FINANCIAL INFRASTRUCTURE
YOU CAN BANK ON

EVERY DAY SOMEWHERE IN THE WORLD…

Yeah! I made a nice model!

Great, let's use it!

Uhhh, ok, can you send me the data?!

No, can I not use it myself? Let's make a tool!

****sweats profusely****

WHY IS THIS SUCH A PAIN?

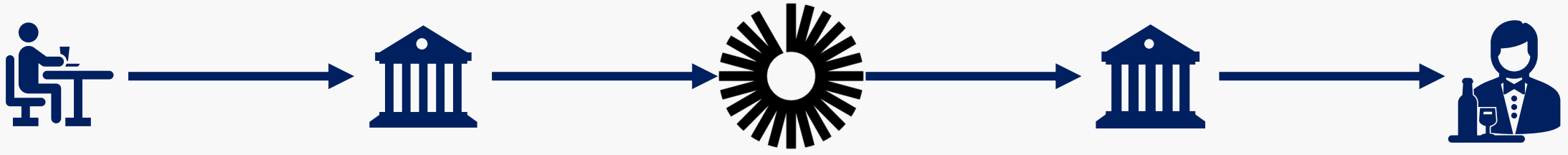
OR people are not familiar with software development, especially for:

- Tests
- Deployment

Software developers often do not understand the intricacies of optimization models and solvers

A lot of things change in the early phases of development

IS THERE ANOTHER WAY? LET'S LOOK AT OUR CASE...



Banking Circle handles payments to provide the "plumbing" of the financial system

As a bank, we want (and are legally required) to screen payments against a list of sanctioned entities. If a part of the payment references a sanctioned entity, it needs to be stopped.

Sounds easy enough? It is string matching, but it is not trivial as we'll see…

A REAL EXAMPLE

OFAC
Office of Foreign Assets Control

Sanctions List Search

Specially Designated Nationals and Blocked Persons list ("SDN List") and all other sanctions lists administered by OFAC, including the Fo
Sanctions Act List, the Sectoral Sanctions Identifications List, the List of Foreign Financial Institutions Subject to Correspondent Account o
Palestinian Legislative Council List. Given the number of lists that now reside in the Sanctions List Search tool, it is strongly recommendec
associated with each returned record. These program codes indicate how a true hit on a returned value should be treated. The Sanctions
identify possible matches between word or character strings as entered into Sanctions List Search, and any name or name component as
sanctions lists. Sanctions List Search has a slider-bar that may be used to set a threshold (i.e., a confidence rating) for the closeness of ar
Sanctions List Search will detect certain misspellings or other incorrectly entered text, and will return near, or proximate, matches, based c
OFAC does not provide recommendations with regard to the appropriateness of any specific confidence rating. Sanctions List Search is ot
and/or the various other sanctions lists; use of Sanctions List Search is not a substitute for undertaking appropriate due diligence. The use
civil liability for any act undertaken as a result of, or in reliance on, such use.

Download the SDN List

Sanctions List Search: Rules for use

Download the Consolidated Non-SDN List

Details:

Type:	Entity	**List:**	SDN
Entity Name:	NAFTIRAN INTERTRADE CO. (NICO) LIMITED	**Program:**	IRAN
		Remarks:	all offices INTERN

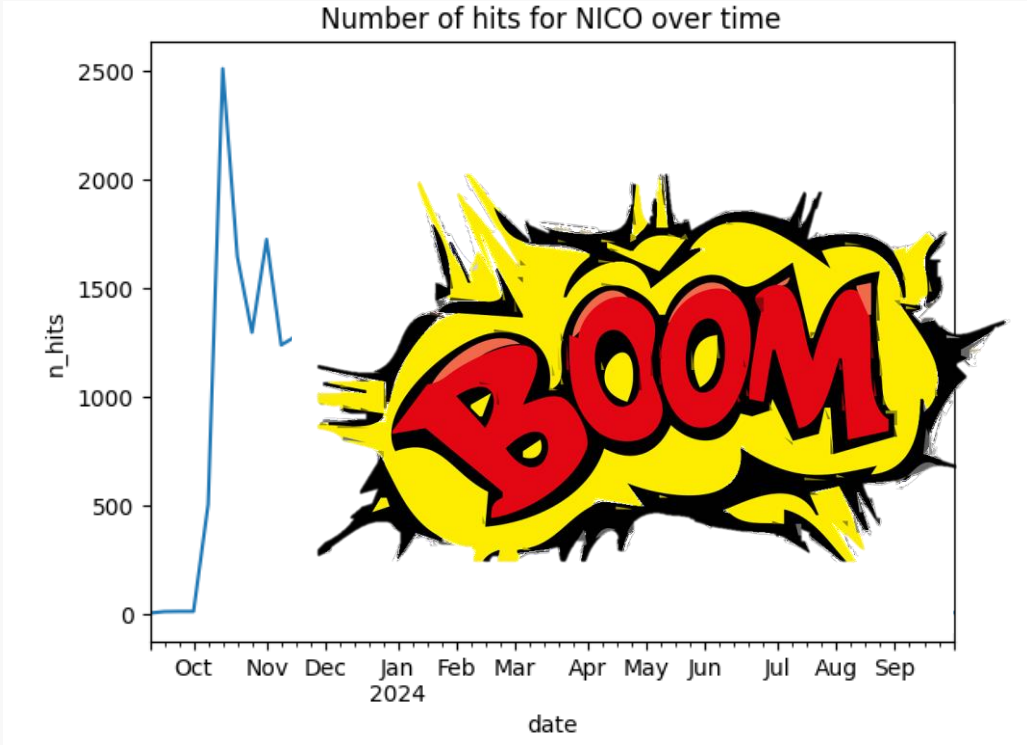
Identifications:

Type	ID#	Country	Issue Date
Additional Sanctions Information -	Subject to Secondary Sanctions		

Aliases:

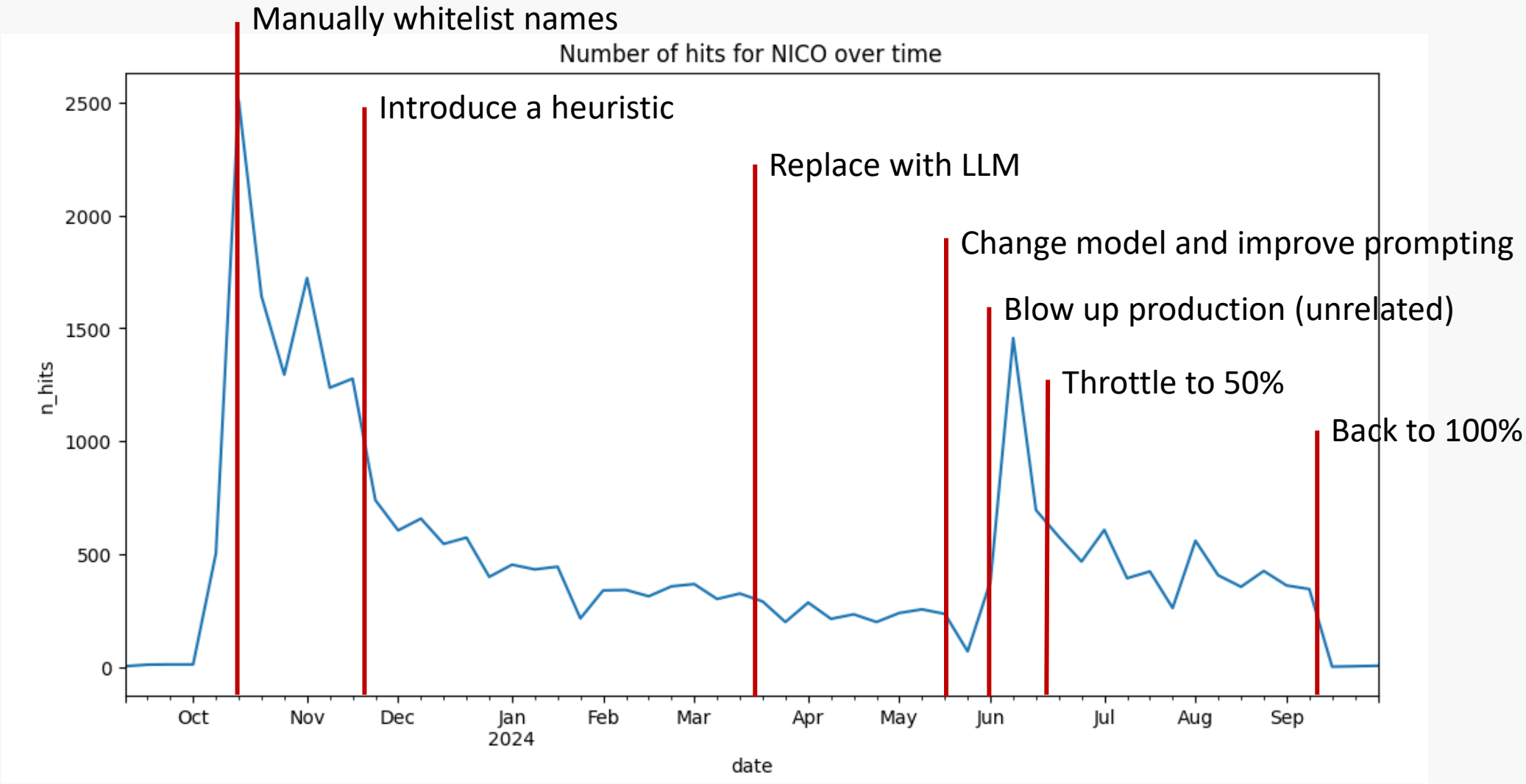
Type	Category	Name
a.k.a.	strong	NICO
a.k.a.	strong	NAFT IRAN INTERTRADE COMPANY LTD
a.k.a.	strong	NAFTIRAN INTERTRADE COMPANY (NICO)
a.k.a.	strong	NAFTIRAN INTERTRADE COMPANY LTD

```
if 'NICO' in creditor.name:
    stop_payment()
```



https://sanctionssearch.ofac.treas.gov/Details.aspx?id=11318

WHAT DID WE DO?

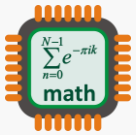


HOW DID WE DO IT?

Business side

Clearly defined and valuable business need

Analytics team owns the application, from infrastructure to code and deployment



Data Scientists are forced to do maths as well as coding, since we own the tools

Tech side

Modular design separating LLM execution from core application

Free And Open-Source (FOSS) models to minimize time to deployment

Simple approach (proompting) before more complex ones (fine-tuning)

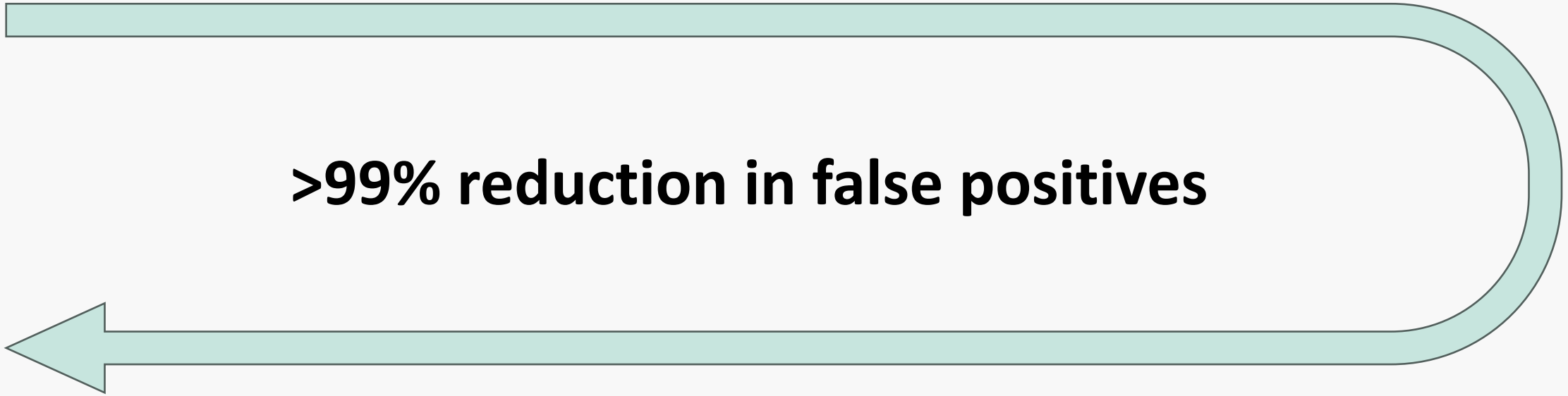
Comprehensive test set that runs separately from regular unit tests

YOU WANT MORE DETAILS? OK…

Can ChatGPT solve this problem?

Ok, but how do we deploy? Should we use Azure directly?

Heard of ollama as a model hosting tool



>99% reduction in false positives

Once first version was running, a lot of experiments on what the best prompt plus model (zephyr) is for now

Converting the Jupyter notebook into prod code: it needs isolation because it is unproved tech

Initial setup was very quick and dirty in a Jupyter notebook. We created the first test file, used mistral as it was popular

WHAT CAN YOU DO?

Force OR professionals to code production code

Be very diligent about testing. This is hard for optimization models, but it can be done. Talk to me if you want to, it'll only take 5-6 hours ☺

FOSS software is essential for quick development speed. Commercial solvers in the OR world are excellent, but buying them can kill your momentum.