

Large Language Models: A new tool in the OR toolbox?

15/10/2024 – Michael Lindahl

ChatGPT is bullshit

Original Paper | [Open access](#) | Published: 08 June 2024

Volume 26, article number 38, (2024) [Cite this article](#)

[Download PDF](#) ↓

✓ You have full access to this [open access article](#)

[Michael Townsen Hicks](#) ✉, [James Humphries](#) & [Joe Slat](#)

Applications of these systems have been plagued by persistent inaccuracies in their output; these are often called “AI hallucinations”. We argue that these falsehoods, and the overall activity of large language models, is better understood as bullshit in the sense explored by Frankfurt (On Bullshit, Princeton, 2005)

All models are wrong, but some are useful
- George E.P. Box

Why

Many OR projects fails because

- Low data quality
- Lack of trust in optimized solutions
- Hard to use decision tools

Can Large Language Models help with?

- Faster development cycles
- Increased trust in the solutions
- Higher-quality solutions, with users in control

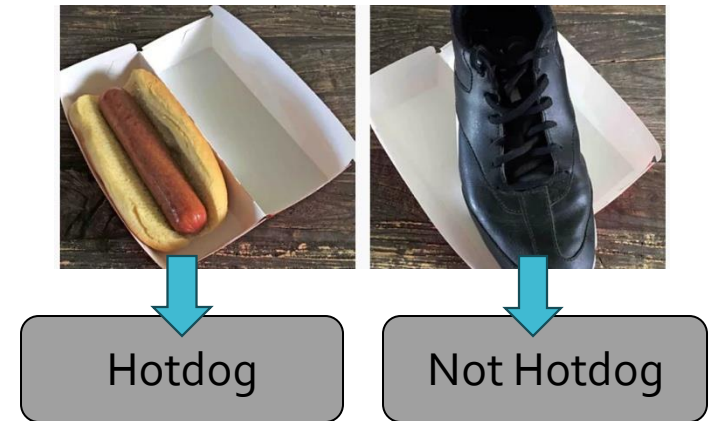
Agenda

- Generative AI and OR
 - What capabilities can be used
 - Case studies
 - Risks and future prospects
-
- Discussion

Generative AI

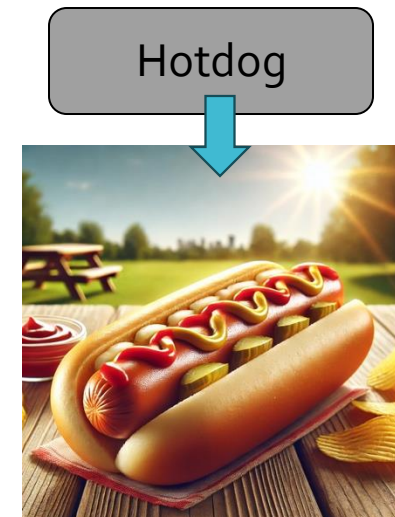
Traditional machine learning

- Identifying patterns, classifying and predicting
- Deep learning, naïve bayes, decision trees, regression,...

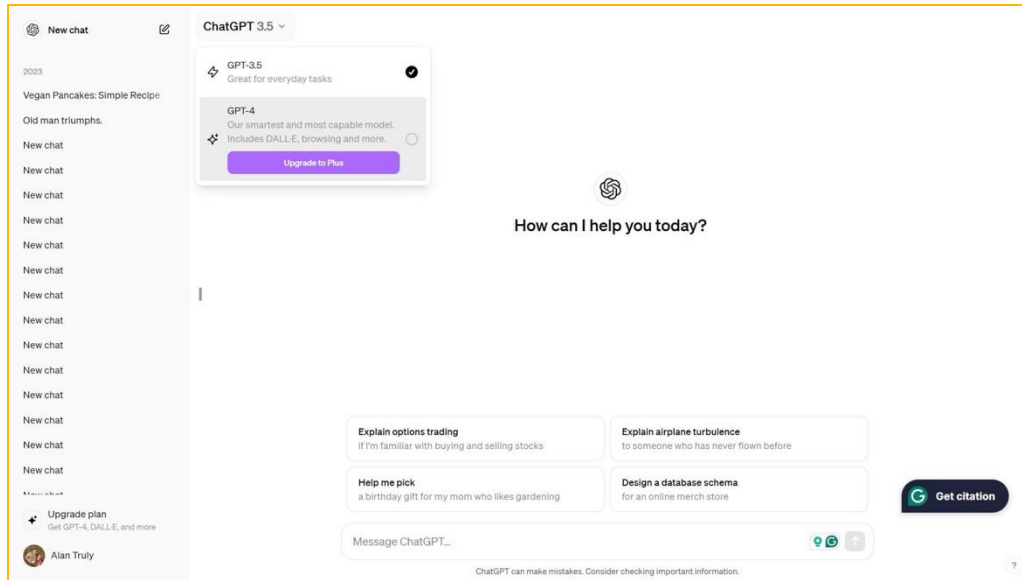


Generative AI

- Generative AI uses machine learning to generate new content
- Is seen in
 - Text generation
 - Image generation
 - Music creation



Large Language Models (LLM's)



- Write text one “word” at a time, based on previous context (GPT)
Lisbon is the capital of Portugal
- Doesn't align with how we want to use a chat
In which country is Lisbon the capital?
In which country is Copenhagen the capital?
- Aligned to human preferences with (ChatGPT)
In which country is Lisbon the capital? Portugal
- OpenAI released ChatGPT in November 2022.
- Reached 100M users after 2 months

The API's opens up possibilities

	<u>Functionality</u>	<u>Example use cases</u>
Templates	Use prompts within code. Use multiple prompts sequentially.	<ul style="list-style-type: none">• Summarize text• Chain multiple templates
Semantic search	Measure how similar two pieces of text are.	<ul style="list-style-type: none">• Find relevant information in a large text corpus.• Classifying
Structured output	Get your data in structured format (e.g. csv or json)	<ul style="list-style-type: none">• Convert an image to text• Extract information from unstructured text
Tools	Have the model decide if it should use a "tool" to solve a specific task.	<ul style="list-style-type: none">• Use a calculator to solve math• Search the web for new information• Write and execute Python code to analyze data

...

See more: <https://platform.openai.com/docs/overview>

LLM's vs. Optimization Models

Large language
Models

Input

- + High flexibility**
 - Can read unstructured data in multiple formats such as text, table, charts and images.
 - Can work with incomplete data

Output

- ÷ Low reliability**
 - Non-factual
 - No clear quality metric
 - Doesn't necessarily obey the constraints it has been given

Mathematical
Optimization Models

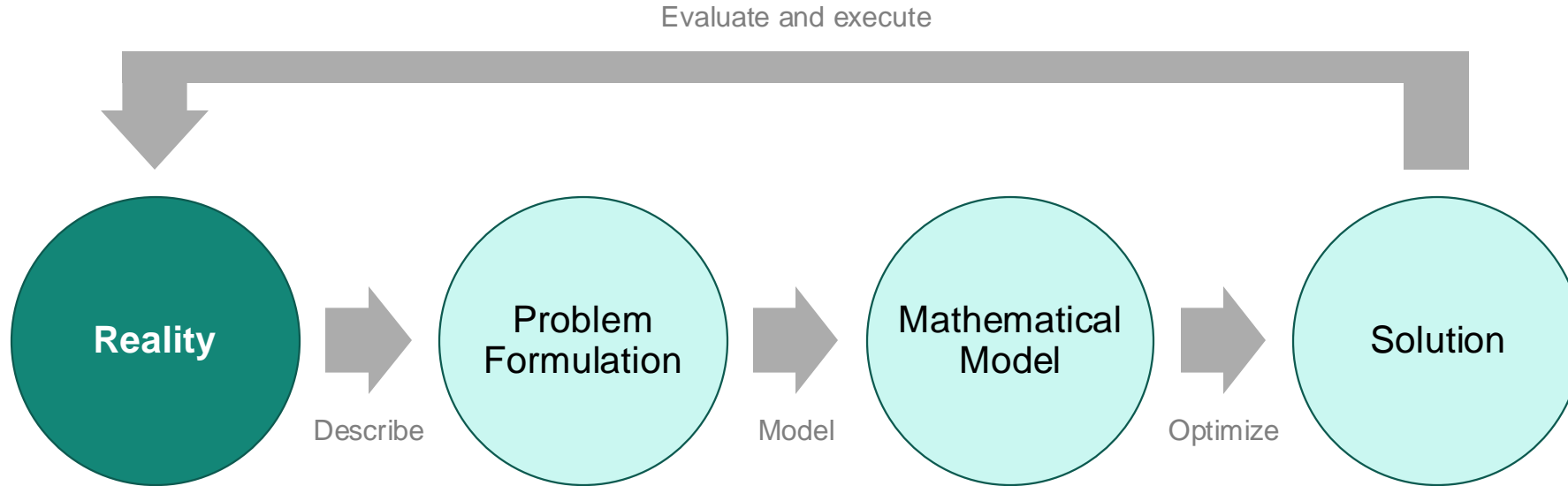
- ÷ Low flexibility**
 - Requires both structured and complete data
 - Requires clear definitions of what everything is.

- + High Reliability**
 - The quality of the output can be aligned with bussiness objectives
 - Interpretable results and relationship
 - Optimality guarantees can be given

Can we get the best of both worlds?

The OR process is ripe for LLM's

OR Process



Suited for Language models

- Mix of unstructured and structured data from interviews, documents, ERP systems etc.
- Incomplete data, context often required

Suited for Mathematical models

- Structured and complete data
- Complex and well defined problems
- Clear definitions and KPI's

Getting to a structured problem formulation

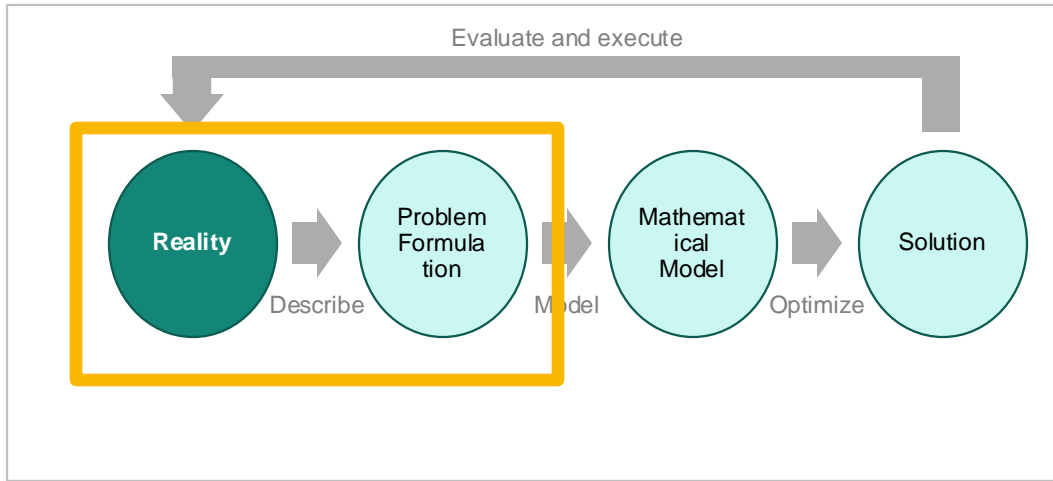


Image data tables to structured data

First Name	Last Name	City	Region	Profession
Peter Benjamin	Parker	Denver	Intermountain	Spideeman
Robin Locksley	Hood	Sherwood Forest	European	Archer
Jean Luc	Picard	La Barre	European	Starship Captain
Jason Charles	Bourne	Nixa	Central	CIA Agent
Clark Joseph	Kent	Smallville	Central	Superman

```
Name, Age, Country, Email, Subscription
Eliana Knox, 28, Canada, eliana.knox@example.com, True
Marco Fischer, 35, Germany, marco.fischer@example.com, False
Aarav Patel, 42, India, aarav.patel@example.com, True
Leyla Wong, 22, Singapore, leyla.wong@example.com, False
Samuel Adams, 31, USA, samuel.adams@example.com, True
Maya Singh, 27, India, maya.singh@example.com, False
Lucas Martin, 19, France, lucas.martin@example.com, True
Olivia Brown, 45, Australia, olivia.brown@example.com, False
Noah Lee, 34, South Korea, noah.lee@example.com, True
Emma López, 30, Spain, emma.lopez@example.com, False
Liam Smith, 24, United Kingdom, liam.smith@example.com, True
Sophia Davis, 50, USA, sophia.davis@example.com, False
Mason Johnson, 29, USA, mason.johnson@example.com, True
Isabella Wilson, 33, USA, isabella.wilson@example.com, False
Jacob Miller, 26, USA, jacob.miller@example.com, True
Ava Taylor, 21, USA, ava.taylor@example.com, False
William Anderson, 37, USA, william.anderson@example.com, True
Mia Thomas, 39, USA, mia.thomas@example.com, False
Ethan Jackson, 41, USA, ethan.jackson@example.com, True
Hanna White, 35, USA, hanna.white@example.com, False
```

Unstructured data from note fields to structured

This customer should be visited after 8:00 and before 13:00

Timewindow = [8:00;13:00]

Not before 14:00

Timewindow = [14:00;17:00]

Scan large documents for relevant info

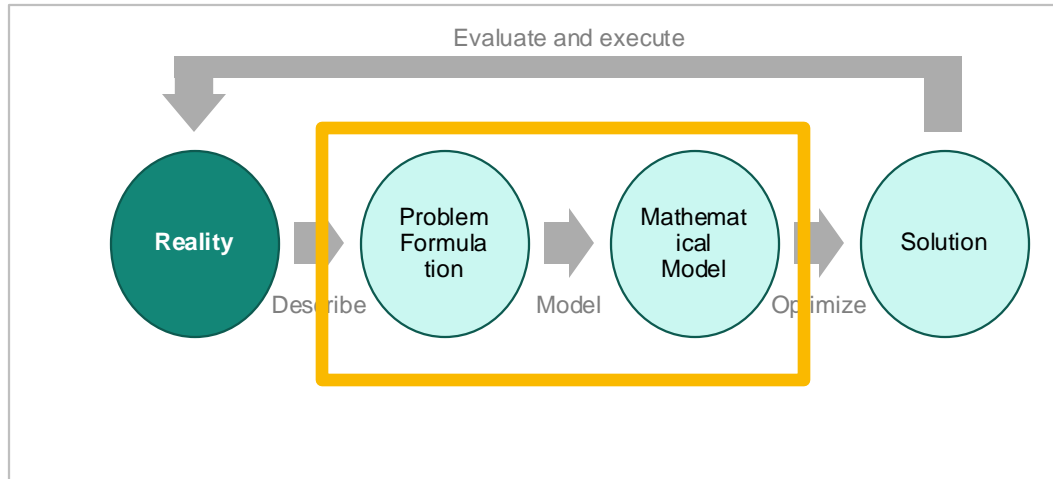


the new mechanism in MkIV overbars, and overstrike. highlighter, which after all is variance, a line is called a rule. n using is called a bar. I'll what they are.)

the line should go in the foreground or the background. Obviously, I choose

Highlighting texts feels good. You can draw attention of people to a **word** or perhaps **even to a whole sentence that spans across multiple lines in such a way that hyper-nation etc. are not affected.**

From problem to mathematical model



V1 of a base model

Work scheduling

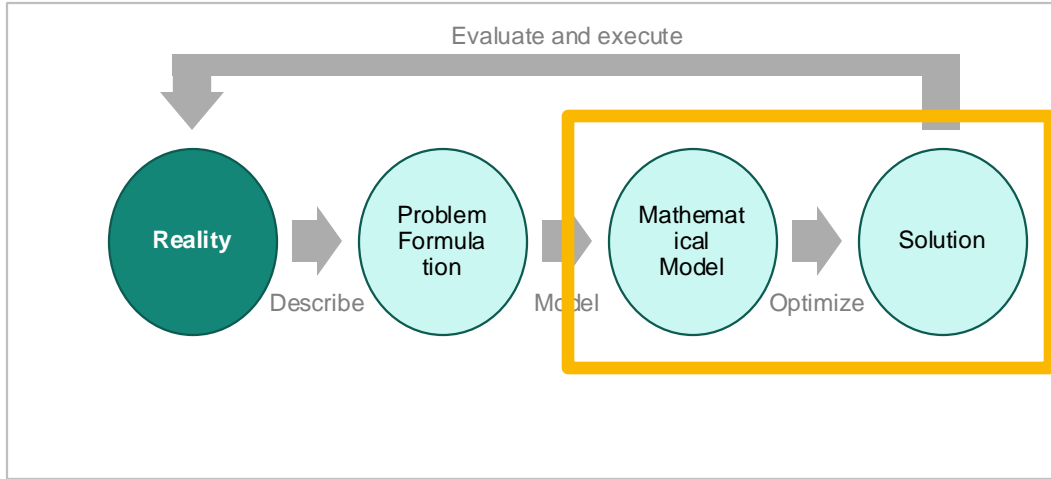
$$\begin{aligned} \min T: & \sum_{i=1}^I \sum_{j=1}^J t y_{ij} && (4.1a) \\ \text{s.t. } & B_i = \sum_{j=1}^J x_{ij} \forall i \in I && (4.1b) \\ & S_j = \sum_{i=1}^I x_{ij} \forall j \in J && (4.1c) \\ & x_{ij} \leq y_{ij} * \min(B_i, S_j) \forall i \in I, \forall j \in J && (4.1d) \\ & y_{ij} \in [0,1] \forall i \in I, \forall j \in J && (4.1e) \\ & x_{ij} \in \mathbb{R} \forall i \in I, \forall j \in J && (4.1f) \end{aligned}$$

Empower users to add side constraints

Bob and Eve can't work at the same time

$$x_1 + x_2 \leq 1$$

Optimization?



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

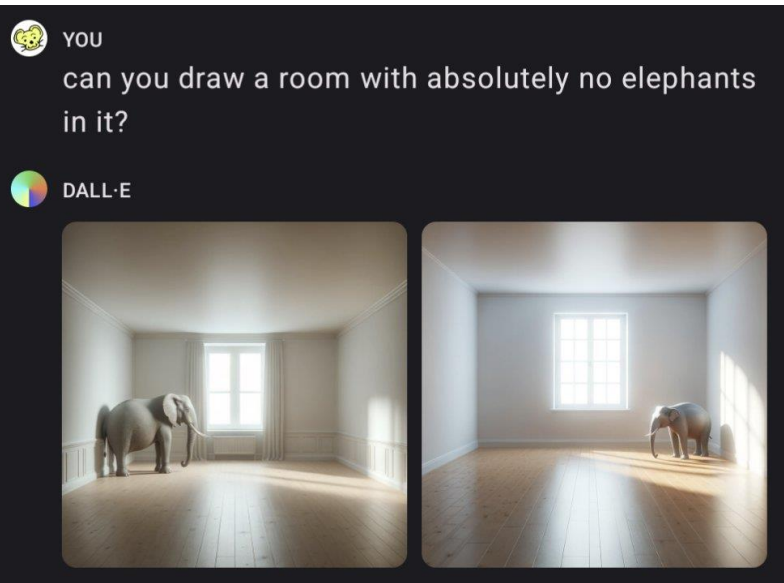
Abstract

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

“we found no evidence of formal reasoning in language models ... Their behavior is better explained by sophisticated pattern matching

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.

No!



LLM's can't reason

GSM-NoOp

Oliver picks 44 kiwis on Friday. Then he picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, but five of them were a bit smaller than average. How many kiwis does Oliver have?

o1-mini: Oliver picks kiwis over three days as follows:

- Friday: 44 kiwis
- Saturday: 58 kiwis
- Sunday: Double the number he picked on Friday, which is $2 \times 44 = 88$ kiwis

However, on Sunday, 5 of these kiwis were smaller than average. We need to subtract them from the Sunday total: 88 (Sunday's kiwis) - 5 (smaller kiwis) = 83 kiwis Now, summing up the kiwis from all three days: 44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis Oliver has a total of 185 kiwis.

Llama3-8B: Let's break it down step by step:

Oliver picks 44 kiwis on Friday. He picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, which is $2 \times 44 = 88$ kiwis.

Five of the kiwis picked on Sunday are a bit smaller than average, so we subtract 5 from the total number of kiwis picked on Sunday: $88 - 5 = 83$ kiwis. Now, let's add up the total number of kiwis Oliver has:

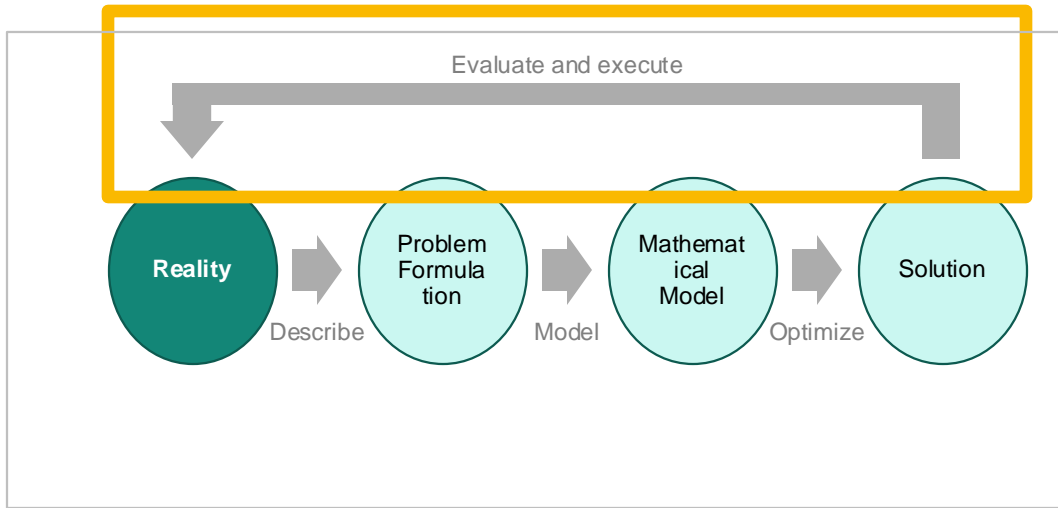
44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis

So, Oliver has 185 kiwis in total.

""Adding a single clause that seems relevant to the question causes significant performance drops (up to 65%) across all state-of-the-art models, even though the clause doesn't contribute to the reasoning chain needed for the final answer.""

"" [...] so fragile, in fact, that changing names can alter results by ~10%!""

Evaluating and turning solutions into actions



Explain solutions

$Obj = 1200$
 $X_1=1, X_2=0$
 $y_{overtime} = 1$

The solution costs 1200, and suggest that Bob works on Monday but will work overtime

Explain infeasibilities

IIS:
 $Sum x \geq 40$
 $X_{Monday} = 0$
 $X_{weekends} = 0$

Bob can't work at least 40 hours when he can't work Monday and weekends.

Compare solutions

Solution 1

Solution 2

Solution 1 is 100 cheaper, but require more employees to work overtime

Case studies

1. LLM's in the backend
2. LLM's as the interface
3. LLM's in OR

Case study: LLM in the backend - Gift finder

FIND GAVEN
Lad os begynde. Hvem skal have en gave?

En mand på 37 år, der hedder Michael. Der elsker teknologi og AI

Ham Hende Barn Teenager Ung Voksen Senior

FIND GAVEIDÉER

Her er nogle gaver, der passer godt til tech-entusiasten Michael

Alle gavekategorier Intet budget

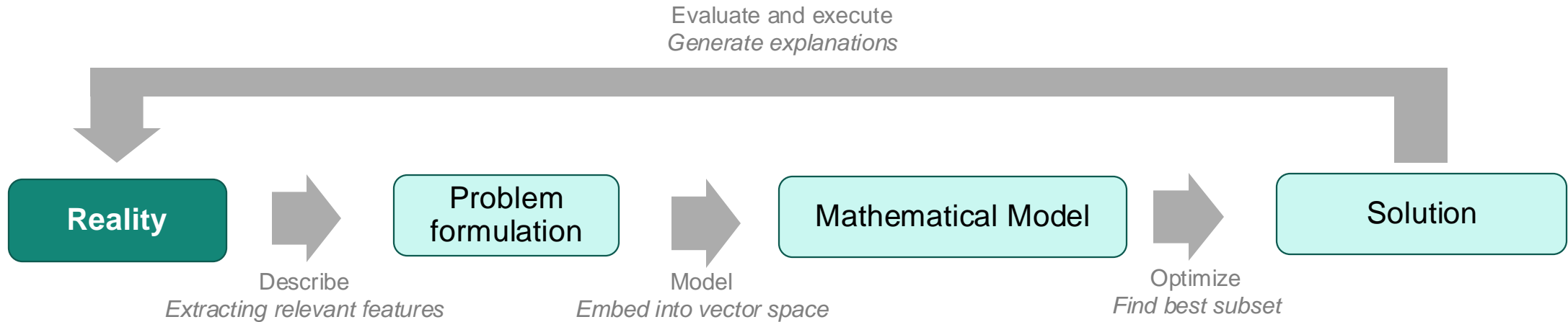
Elon Musk // Walter Isaacson
ELON MUSK
For selvom Michael allerede er dybt engageret i teknologi og AI, vil denne biografi om Elon Musk kunne inspirere ham yderligere og give ham et indblik i en af de mest fascinerende og kontroversielle
300 kr
SE HOS BOG-IDE

DJI MINI 3 - DRONE - DRONE M. KAMERA
4.5
Giv Michael en DJI Mini 3 drone med kamera, så han kan udforske sin passion for teknologi og AI fra en helt ny vinkel. Denne skønne drone vil give ham mulighed for at skabe professionelle filmoptagelser og udforske den teknologiske verden fra
3.399 kr
SE HOS FØNIKS COMPUTER

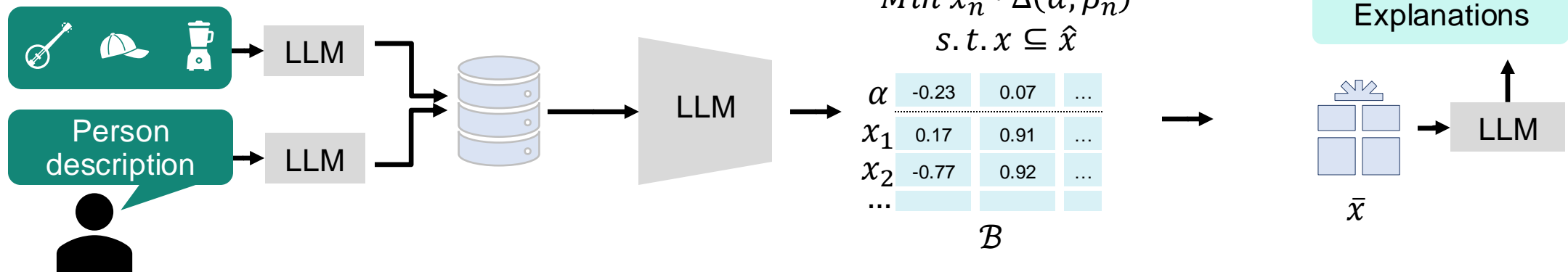
10283 LEGO NASA SPACE
4.6
GRATIS FRAGT
Den fantastiske LEGO Creator Expert NASA-rumfærgen Discovery (10283) vil være den perfekte gave til Michael, da han elsker teknologi og AI. Med dette imponerende byggesæt kan han dykke ned i rummets vidundere og
1.550 kr
SE HOS BOG-IDE

1. Users give a description of the recipient of the gift
2. The app searches through products from 20 webshops
3. Give the user the best matches with an explanation of how they match.

Case study: Gift finder



Case study: Finding the “optimal” gift:



Case study: LLM as the Interface

`https://Database.build`

I am building an application to help me route my trucks to visit my customers.... I have a list with all my customers, a customer have a name and phone number, and can also have a time window. I also have a list of trucks. I should then be able to save routes for my trucks how they visit my customers.

Case study: LLM for OR - GurobiGPT

- I am visiting Porto and want to taste some good port wine and therefore decided to visit some wineries. I have 6 hours, and want to maximize the ratings of the ones I visit.
I want to visit at least one in top-3 biggest wineries..

DATA:

Cellar Name, Tasting Duration (hours), Prestige Rating (1-10), Top-3 Largest?

Graham's, 2, 10, Yes

Sandeman, 1.5, 8, Yes

Ferreira, 1, 7, No

Offley, 1.5, 6, No

Taylor's, 2, 9, Yes

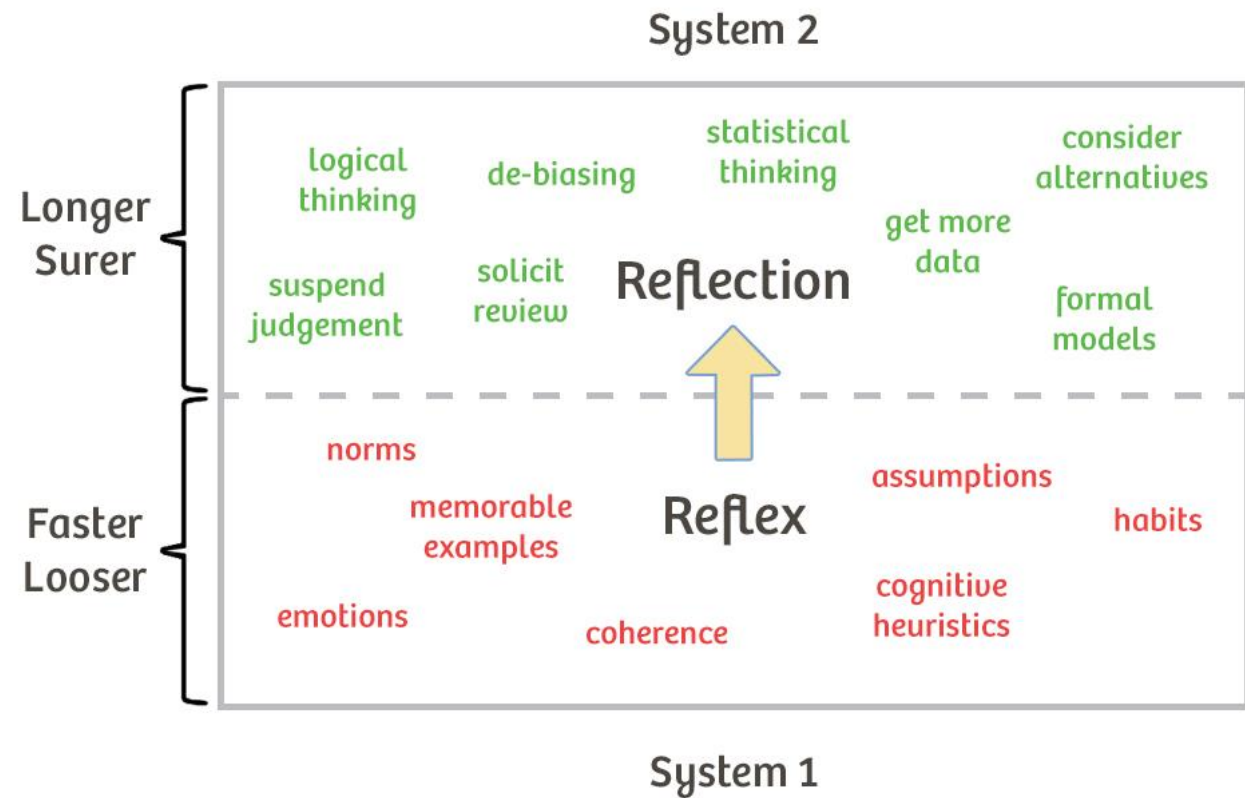
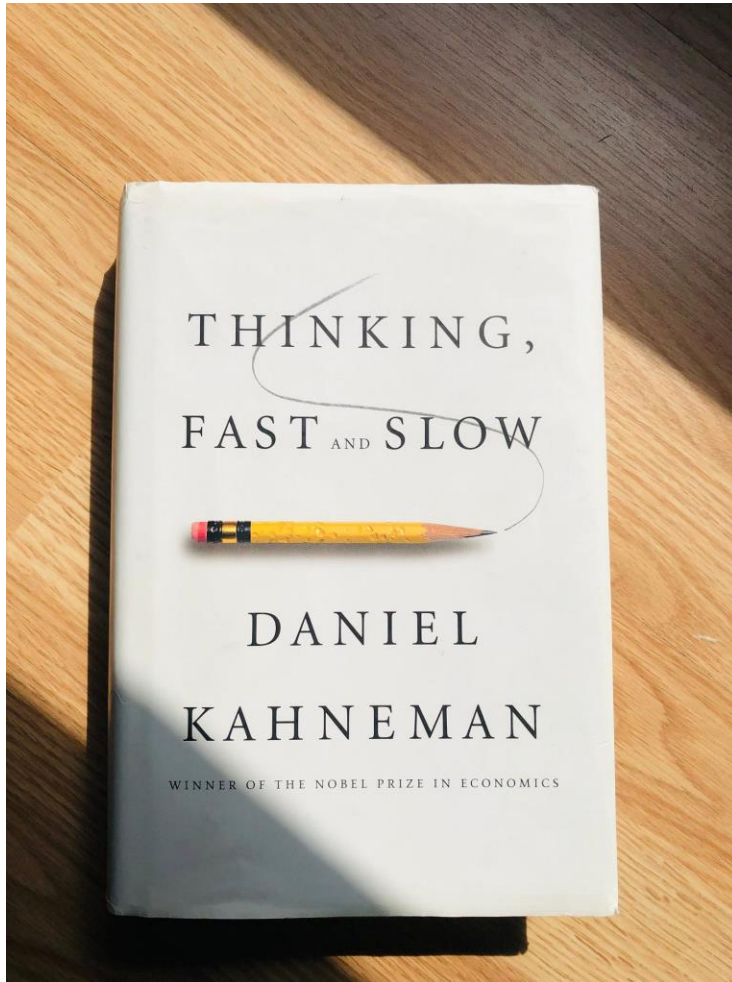
Croft, 1.2, 7, No

Which of these should i visit?

Risks and Future prospects

- Improved quality and performance of LLM's
- More best practices and better tooling
- Keeping sensitive data to yourself
- Lower costs

Thinking, Fast and Slow



Any questions?

Risk, enabler or distraction?

Let's discuss!

Questions:

1. Where do you see generative AI working?
2. Where do you see it not working?
3. How do you see generative AI work in the context of OR?
4. Is it mostly an disruptor, enabler or distraction?

Last remarks!

- Start small
- Define input and output
- Keep human in the loop
- Get started!

Meget problemløsning kræver to domæner



Virkelig verden

Min kæreste og jeg skal have en sort kop kaffe hver

Det bliver 54kr. ialt

Godt til LLM'er

- Ustruktureret data
- Context er vigtigt

Model verden

Bestilling: 2 Café Americano

$2 * 27kr. = 54kr.$

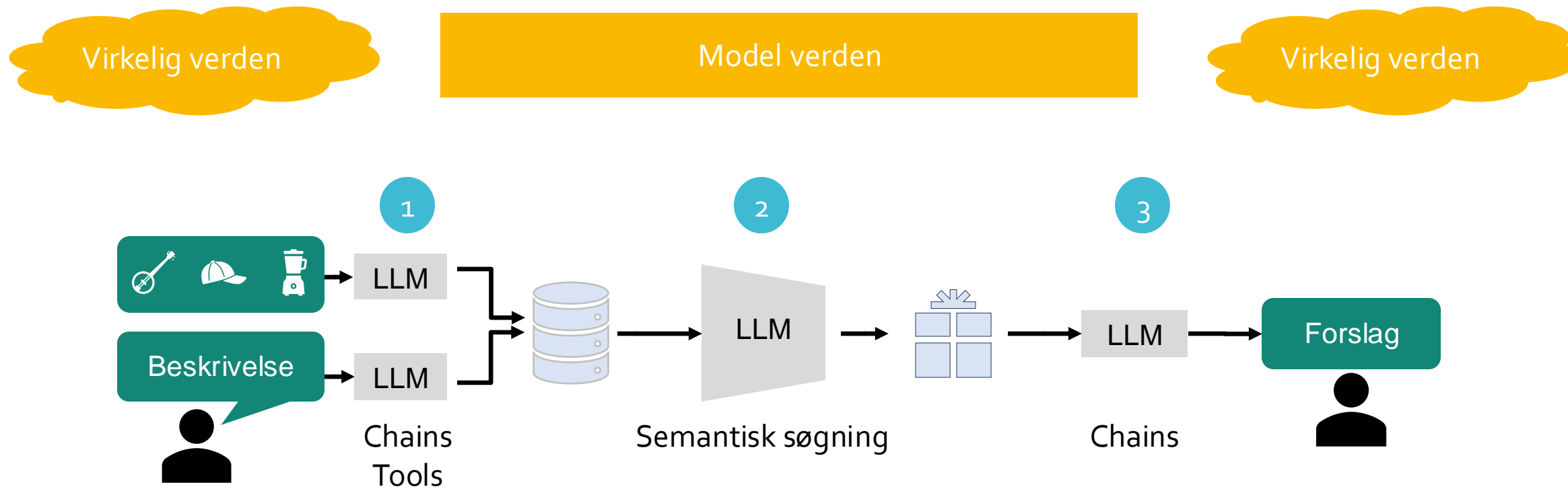
Pris: 54kr.

Godt til traditionel software

- Struktureret og komplet data
- Komplekse problemer

Case study: FindGaven.AI

3 steder med sammenlagt 12 LLM funktioner



1

- Samle data fra 16 forskellige webshops til struktureret data
- Detekter alder og andet relevant info fra brugerbeskrivelse

2

- Semantisk søgning for at finde gaver der passer til beskrivelsen

3

- Chains for at generer forklaringer til brugeren