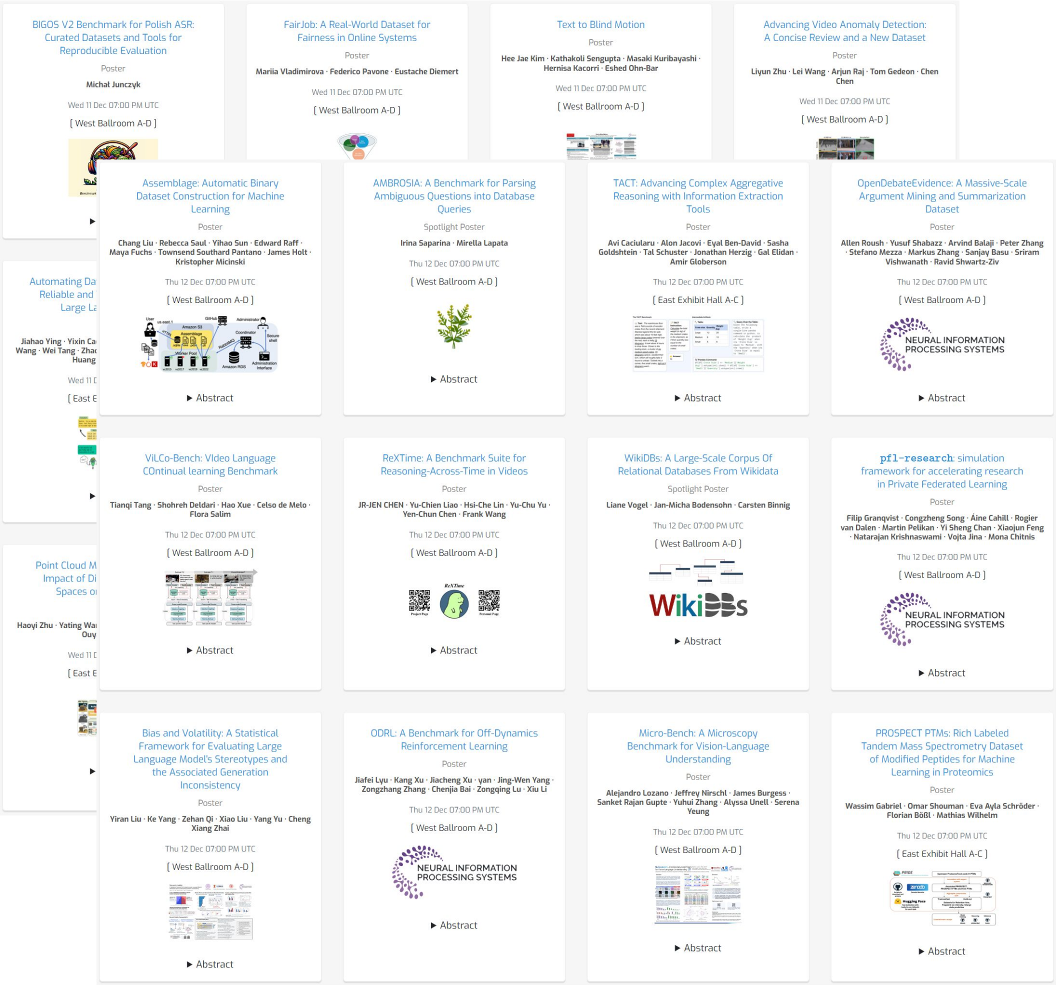
The 39th Annual AAAI Conference
on Artificial Intelligence
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, PENNSYLVANIA, USA

Evaluating LLM Reasoning in the Operations Research Domain with ORQA

Mahdi Mostajabdaveh, Timothy Tin Long Yu, Samarendra Chandan Bindu Dash, Rindra Ramamonjison,
Jabo Serge Byusa, Giuseppe Carenini, Zirui Zhou, Yong Zhang

UBC
HUAWEI
UNIVERSITY OF TORONTO

Another Benchmark??



BIGOS V2 Benchmark for Polish ASR: Curated Datasets and Tools for Reproducible Evaluation
Poster
Michał Junczyk
Wed 11 Dec 07:00 PM UTC
[West Ballroom A-D]

FairJob: A Real-World Dataset for Fairness in Online Systems
Poster
Mariia Vladimirova · Federico Pavone · Eustache Diemert
Wed 11 Dec 07:00 PM UTC
[West Ballroom A-D]

Text to Blind Motion
Poster
Hee Jae Kim · Kathakoli Sengupta · Masaki Kuribayashi · Hernisa Kacorri · Eshed Ohn-Bar
Wed 11 Dec 07:00 PM UTC
[West Ballroom A-D]

Advancing Video Anomaly Detection: A Concise Review and a New Dataset
Poster
Liyun Zhu · Lei Wang · Arjun Raj · Tom Gedeon · Chen Chen
Wed 11 Dec 07:00 PM UTC
[West Ballroom A-D]

Automating Da
Reliable and
Large La
Jiahao Ying · Yixin Ca
Wang · Wei Tang · Zhao
Huang
Wed 11 D
[East E

Point Cloud M
Impact of Di
Spaces o
Haoyi Zhu · Yating Wa
Ouy
Wed 11 D
[East E

Assemblage: Automatic Binary Dataset Construction for Machine Learning
Poster
Chang Liu · Rebecca Saul · Yihao Sun · Edward Raff · Maya Fuchs · Townsend Southard Pantano · James Holt · Kristopher Micinski
Thu 12 Dec 07:00 PM UTC
[West Ballroom A-D]
▸ Abstract

AMBROSIA: A Benchmark for Parsing Ambiguous Questions into Database Queries
Spotlight Poster
Irina Saparina · Mirella Lapata
Thu 12 Dec 07:00 PM UTC
[West Ballroom A-D]
▸ Abstract

TACT: Advancing Complex Aggregative Reasoning with Information Extraction Tools
Poster
Avi Caciularu · Alon Jacovi · Eyal Ben-David · Sasha Goldshtein · Tal Schuster · Jonathan Herzig · Gal Elidan · Amir Globerson
Thu 12 Dec 07:00 PM UTC
[East Exhibit Hall A-C]
▸ Abstract

OpenDebateEvidence: A Massive-Scale Argument Mining and Summarization Dataset
Poster
Allen Roush · Yusuf Shabazz · Arvind Balaji · Peter Zhang · Stefano Mezza · Markus Zhang · Sanjay Basu · Sriram Vishwanath · Ravid Shwartz-Ziv
Thu 12 Dec 07:00 PM UTC
[West Ballroom A-D]
▸ Abstract

ViLCo-Bench: VIdeo Language COntinual learning Benchmark
Poster
Tianqi Tang · Shohreh Deldari · Hao Xue · Celso de Melo · Flora Salim
Thu 12 Dec 07:00 PM UTC
[West Ballroom A-D]
▸ Abstract

ReXTime: A Benchmark Suite for Reasoning-Across-Time in Videos
Poster
JR-JEN CHEN · Yu-Chien Liao · Hsi-Che Lin · Yu-Chu Yu · Yen-Chun Chen · Frank Wang
Thu 12 Dec 07:00 PM UTC
[West Ballroom A-D]
▸ Abstract

WikiDBs: A Large-Scale Corpus Of Relational Databases From Wikidata
Spotlight Poster
Liane Vogel · Jan-Micha Bodensohn · Carsten Binnig
Thu 12 Dec 07:00 PM UTC
[West Ballroom A-D]
▸ Abstract

`pfl-research`: simulation framework for accelerating research in Private Federated Learning
Poster
Filip Granqvist · Congzheng Song · Áine Cahill · Rogier van Dalen · Martin Pelikan · Yi Sheng Chan · Xiaojun Feng · Natarajan Krishnaswami · Vojta Jina · Mona Chitnis
Thu 12 Dec 07:00 PM UTC
[West Ballroom A-D]
▸ Abstract

Bias and Volatility: A Statistical Framework for Evaluating Large Language Model's Stereotypes and the Associated Generation Inconsistency
Poster
Yiran Liu · Ke Yang · Zehan Qi · Xiao Liu · Yang Yu · Cheng Xiang Zhai
Thu 12 Dec 07:00 PM UTC
[West Ballroom A-D]
▸ Abstract

ODRL: A Benchmark for Off-Dynamics Reinforcement Learning
Poster
Jiafei Lyu · Kang Xu · Jiacheng Xu · yan · Jing-Wen Yang · Zongzhang Zhang · Chenjia Bai · Zongqing Lu · Xiu Li
Thu 12 Dec 07:00 PM UTC
[West Ballroom A-D]
▸ Abstract

Micro-Bench: A Microscopy Benchmark for Vision-Language Understanding
Poster
Alejandro Lozano · Jeffrey Nirschl · James Burgess · Sanket Rajan Gupte · Yuhui Zhang · Alyssa Unell · Serena Yeung
Thu 12 Dec 07:00 PM UTC
[West Ballroom A-D]
▸ Abstract

PROSPECT PTMs: Rich Labeled Tandem Mass Spectrometry Dataset of Modified Peptides for Machine Learning in Proteomics
Poster
Wassim Gabriel · Omar Shouman · Eva Ayla Schröder · Florian Bößl · Mathias Wilhelm
Thu 12 Dec 07:00 PM UTC
[East Exhibit Hall A-C]
▸ Abstract

https://neurips.cc/virtual/2024/events/datasets-benchmarks-2024

Why ORQA?

- Expert **hand-crafted** dataset
- **Difficult questions** from a highly specialized technical domain
  - Limited exposure during training
  - Multi-step reasoning
- **Significant gap** between tested LLMs and domain experts

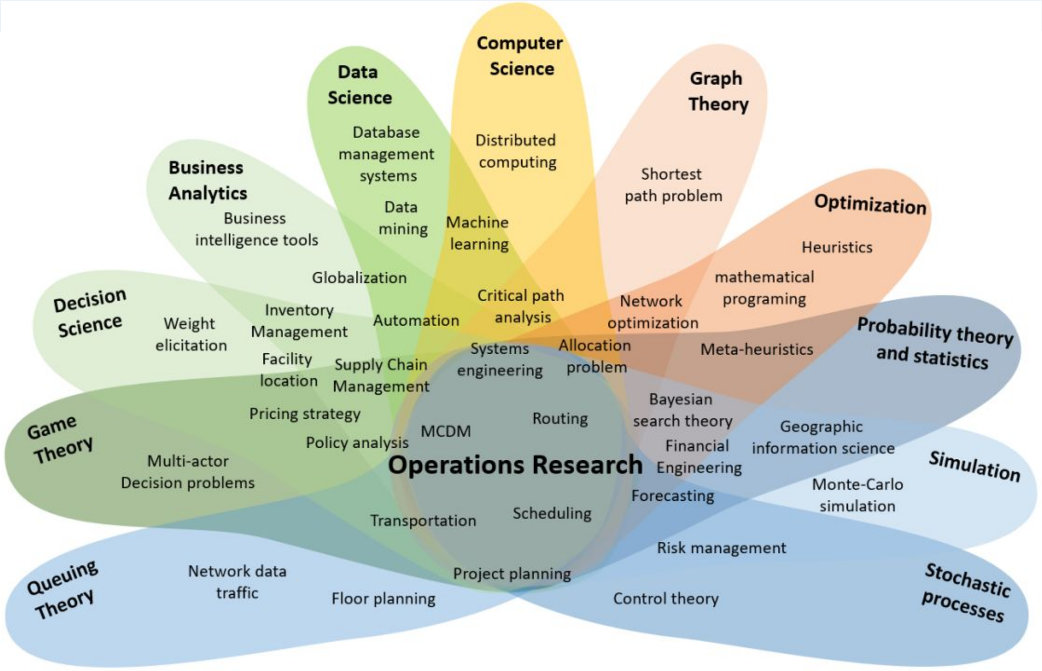
0.77

0.93

Overview

- Background & Motivation
- What is ORQA?
- How we created ORQA
- Benchmarking LLMs
- Findings
- Future works

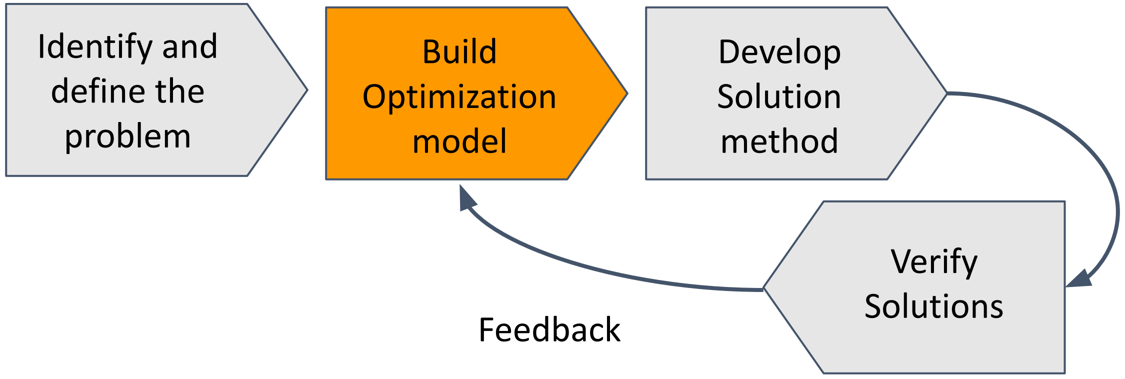
Background



The Importance of Operations Research (OR)

- **Wide-Ranging Applications:** Integral to industries like logistics, manufacturing, healthcare, finance, supply chain management, and urban planning.
- **Critical impact**: Automated decision-making, enhances efficiency, reduces costs.

- **Optimization Modeling bottleneck of OR wide adaptation.**
- **Automating the translation of real-world problems into optimization models can revolutionize how industries use OR.**
  - Lowers cost of OR projects.
  - Democratizes access to expert-level optimization
  - Expands adoption of OR



Identify and define the problem

Build Optimization model

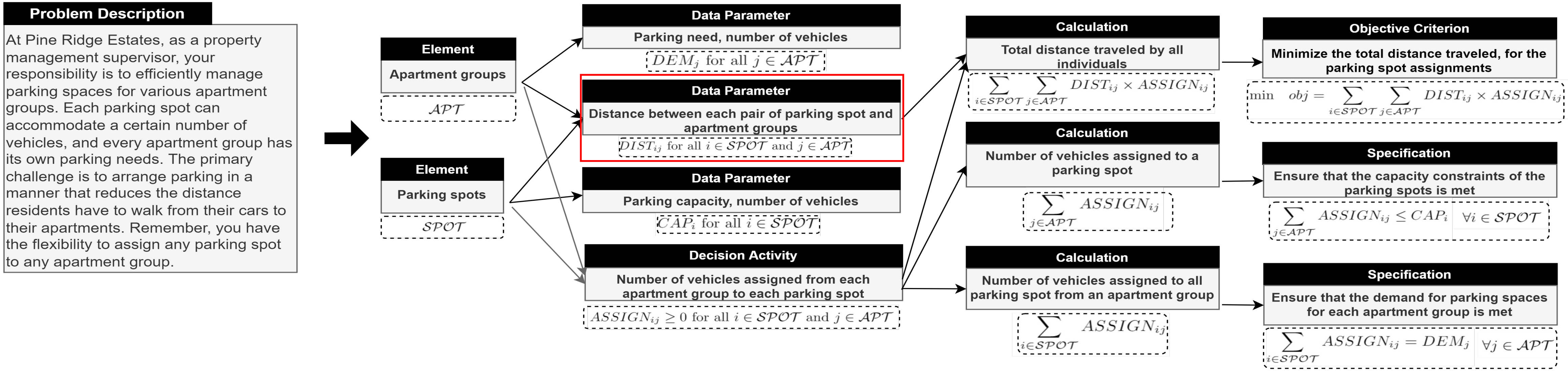
Develop Solution method

Verify Solutions

Feedback

Background - Complexity of Optimization Modeling Tasks

- **Multiple Components:**
  - **Elements:** Main entities in the problem
  - **Decision activities:** Represent choices that impact the model's outcomes.
  - **Data Parameters:** Critical data that are required to inform the model.
  - **Objective Criterion:** The goals to be maximized or minimized, such as cost, time, or profit.
  - **Constraints:** Rules and limitations that bind the model
- **Interdependencies:**
  - A change in one component (e.g., decision variables) directly affects others (e.g., constraints).
- **Implicit Components & Ambiguities:**
  - Not all model components are explicitly stated in problem descriptions and require expert insight to identify.
  - Problem owners often provide vague or ambiguous descriptions due to lack optimization expertise.



Problem Description

At Pine Ridge Estates, as a property management supervisor, your responsibility is to efficiently manage parking spaces for various apartment groups. Each parking spot can accommodate a certain number of vehicles, and every apartment group has its own parking needs. The primary challenge is to arrange parking in a manner that reduces the distance residents have to walk from their cars to their apartments. Remember, you have the flexibility to assign any parking spot to any apartment group.

Element

Apartment groups

$\mathcal{APT}$

Element

Parking spots

$\mathcal{SPOT}$

Data Parameter

Parking need, number of vehicles

DEM_j for all $j \in \mathcal{APT}$

Data Parameter

Distance between each pair of parking spot and apartment groups

$DIST_{ij}$ for all $i \in \mathcal{SPOT}$ and $j \in \mathcal{APT}$

Data Parameter

Parking capacity, number of vehicles

CAP_i for all $i \in \mathcal{SPOT}$

Decision Activity

Number of vehicles assigned from each apartment group to each parking spot

$ASSIGN_{ij} \geq 0$ for all $i \in \mathcal{SPOT}$ and $j \in \mathcal{APT}$

Calculation

Total distance traveled by all individuals

$$\sum_{i \in \mathcal{SPOT}} \sum_{j \in \mathcal{APT}} DIST_{ij} \times ASSIGN_{ij}$$

Calculation

Number of vehicles assigned to a parking spot

$$\sum_{j \in \mathcal{APT}} ASSIGN_{ij}$$

Calculation

Number of vehicles assigned to all parking spot from an apartment group

$$\sum_{i \in \mathcal{SPOT}} ASSIGN_{ij}$$

Objective Criterion

Minimize the total distance traveled, for the parking spot assignments

$$\min \quad obj = \sum_{i \in \mathcal{SPOT}} \sum_{j \in \mathcal{APT}} DIST_{ij} \times ASSIGN_{ij}$$

Specification

Ensure that the capacity constraints of the parking spots is met

$$\sum_{j \in \mathcal{APT}} ASSIGN_{ij} \leq CAP_i \quad \forall i \in \mathcal{SPOT}$$

Specification

Ensure that the demand for parking spaces for each apartment group is met

$$\sum_{i \in \mathcal{SPOT}} ASSIGN_{ij} = DEM_j \quad \forall j \in \mathcal{APT}$$

Motivation

Challenges

- **Complex Reasoning Required:**
  - The interrelated nature of components demands multi-step, expert-level reasoning to build accurate models.
- **Dataset Scarcity:**
  - Operates in an underrepresented field where training data is limited, making it a rigorous testbed for LLM capabilities.
  - There are very few small optimization modeling datasets.

What Makes ORQA Unique?

- Multi-Step Reasoning
- Expert-Level Domain Knowledge
- Complex Interactions
- Niche Technical Domain

Impact

- **Advancing LLM Research:**
  - Provides insights into the limitations and potential improvements for LLMs in specialized domains.
- **Automation in OR:**
  - Paves the way for automating complex decision-making tasks in real-world operations research applications.

ORQA Example

Dataset Instance

Problem description:

You are an operations manager in the agricultural sector. Your task is to streamline the process of getting crops from farms, processing or storing them as needed, and finally distributing them to various markets or direct consumers.
...
You also need to ensure the produce reaches markets and consumers in its best state, meets demands, and ensures a steady supply.

Question:

What is the type of optimization model related to this problem?

Options & answer:

A - Mixed-Integer Linear Programming
B - Linear Programming
C - Integer Linear Programming
D - Non-linear Programming

Expert Reasoning (val. set only)

Reasoning steps:

Step 1: What are the decision activities in the model?
- Amount of product to cultivate, Selecting a transportation mode (truck, rail, etc.)

Step 2: What is the type of values they can get?
- Continuous and integer

Step 3: Is there any Non-linear relationship presented in the problem?
- Constraints and objective function does not define a non-linear relation

Step 4: Given the variable types and linear or non-linear relationship between them, what is the optimization model type?
- There are both continuous and integer variables and there is no non-linear relation exist in the optimization problem. Therefore, the most suitable optimization model for this problem is Mixed-Integer Linear Programming (MILP).

ORQA

Statistics & Characteristics

- 1513 instances
- 1468 for testing; 45 for validation (ICL)
- 20 diverse application domains (e.g., healthcare, urban design, HR, petroleum)
- From each domain:
  - > 3 optimization problems
  - 60 – 90 multiple choice questions
- Standardized optimization model complexity
  - Mathematical model is within pre-defined limits (e.g., number of decision variables)

Component	**μ**	**σ**
Sets	1.97	0.89
Parameters	4.08	2.19
Variables	3.14	2.29
Objectives	1.00	0.00
Constraints	4.60	3.21

ORQA – Question Types

11 Question Types

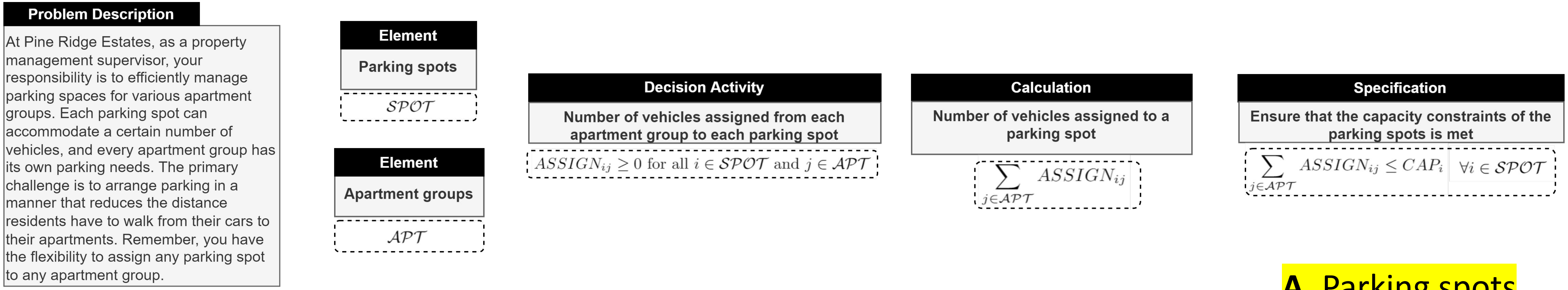
- Understanding high-level problem specifications
    - Objective identification
    - Explicit constraint identification
    - Problem categorization
- Entities identification
    - Optimization type (e.g., linear, non-linear)
    - Set, decision activities, implicit constraint identification
- Relationship between Entities
    - Parameters and variables in objective / constraints
    - Meaning of calculations in objective / constraints
    - Set that a constraint is applied on

ORQA – Question Types

Set that a constraint is applied on

Question: Which of the following system element(s) is the Parking Spot Capacity constraint applied on?

A. Parking spots, **B.** Apartment group, **C.** Parking spots, Apartment group, vehicle count, **D.** Parking spots, Apartment group



Problem Description

At Pine Ridge Estates, as a property management supervisor, your responsibility is to efficiently manage parking spaces for various apartment groups. Each parking spot can accommodate a certain number of vehicles, and every apartment group has its own parking needs. The primary challenge is to arrange parking in a manner that reduces the distance residents have to walk from their cars to their apartments. Remember, you have the flexibility to assign any parking spot to any apartment group.

Element

Parking spots

$\mathcal{SPOT}$

Element

Apartment groups

$\mathcal{APT}$

Decision Activity

Number of vehicles assigned from each apartment group to each parking spot

$ASSIGN_{ij} \geq 0 \text{ for all } i \in \mathcal{SPOT} \text{ and } j \in \mathcal{APT}$

Calculation

Number of vehicles assigned to a parking spot

$\sum_{j \in \mathcal{APT}} ASSIGN_{ij}$

Specification

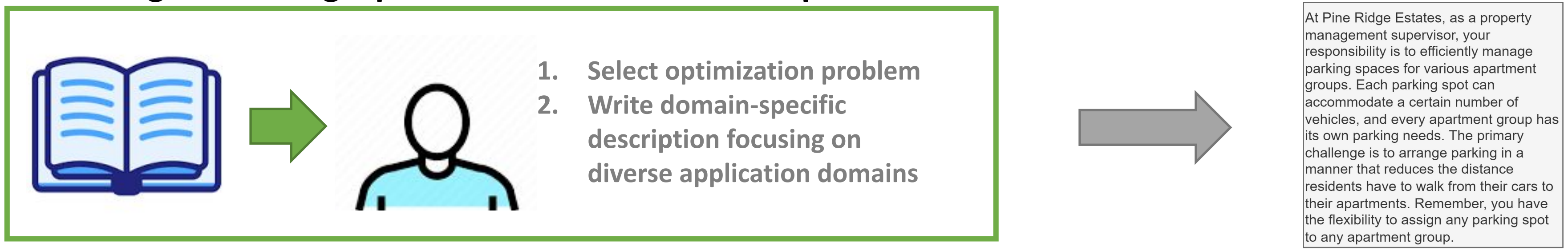
Ensure that the capacity constraints of the parking spots is met

$\sum_{j \in \mathcal{APT}} ASSIGN_{ij} \leq CAP_i \quad \forall i \in \mathcal{SPOT}$

A. Parking spots

Dataset Creation and Verification

Selecting & Creating Optimization Problem Descriptions

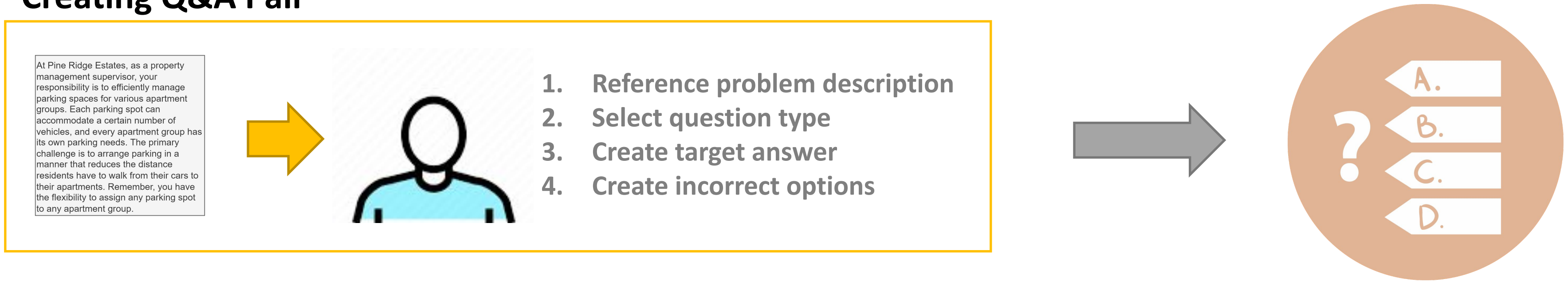


1. Select optimization problem
2. Write domain-specific description focusing on diverse application domains

Problem Description

At Pine Ridge Estates, as a property management supervisor, your responsibility is to efficiently manage parking spaces for various apartment groups. Each parking spot can accommodate a certain number of vehicles, and every apartment group has its own parking needs. The primary challenge is to arrange parking in a manner that reduces the distance residents have to walk from their cars to their apartments. Remember, you have the flexibility to assign any parking spot to any apartment group.

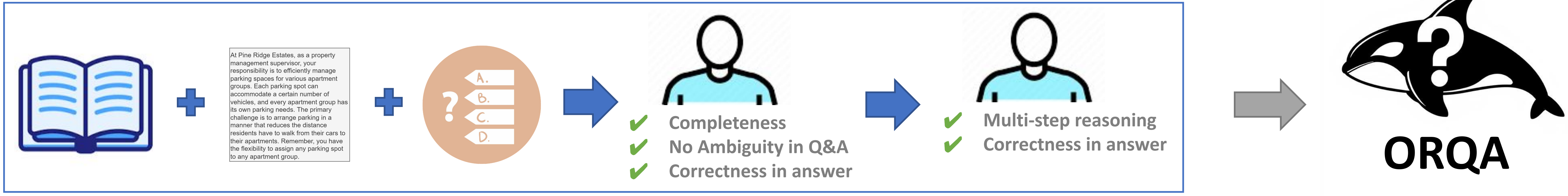
Creating Q&A Pair



At Pine Ridge Estates, as a property management supervisor, your responsibility is to efficiently manage parking spaces for various apartment groups. Each parking spot can accommodate a certain number of vehicles, and every apartment group has its own parking needs. The primary challenge is to arrange parking in a manner that reduces the distance residents have to walk from their cars to their apartments. Remember, you have the flexibility to assign any parking spot to any apartment group.

1. Reference problem description
2. Select question type
3. Create target answer
4. Create incorrect options

Instance Verification



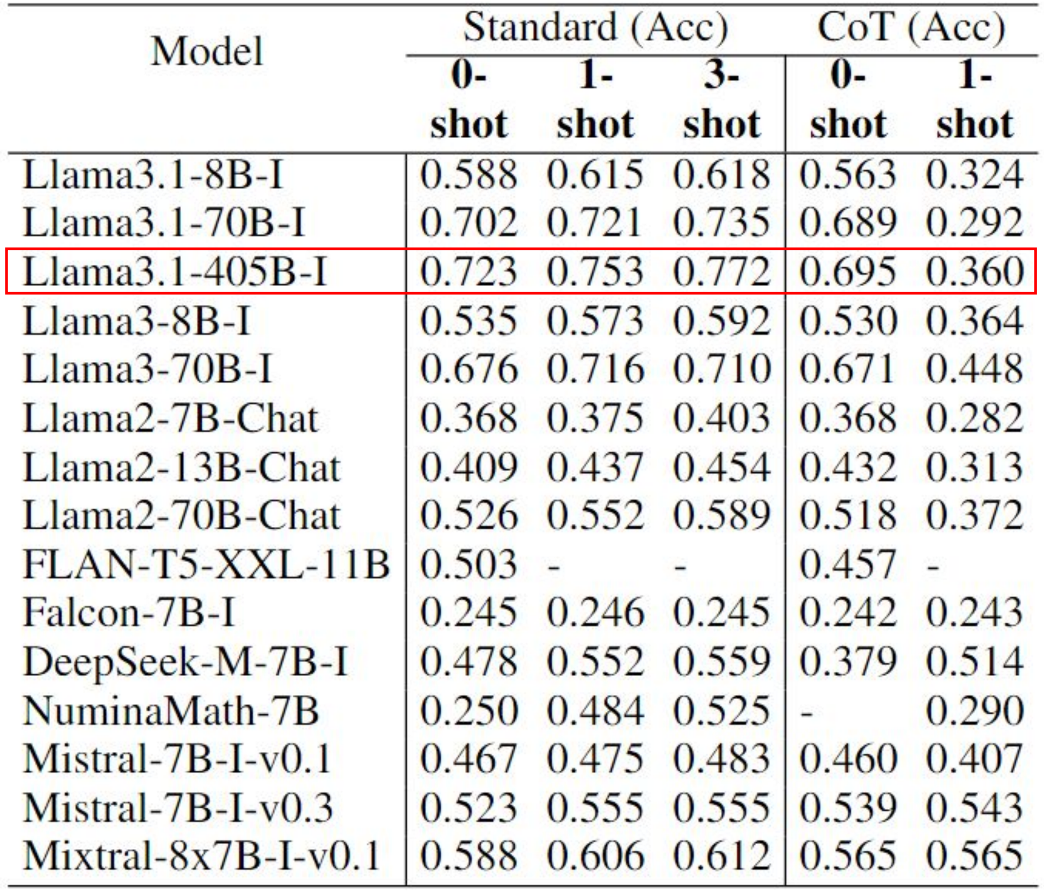
At Pine Ridge Estates, as a property management supervisor, your responsibility is to efficiently manage parking spaces for various apartment groups. Each parking spot can accommodate a certain number of vehicles, and every apartment group has its own parking needs. The primary challenge is to arrange parking in a manner that reduces the distance residents have to walk from their cars to their apartments. Remember, you have the flexibility to assign any parking spot to any apartment group.

- ✓ Completeness
- ✓ No Ambiguity in Q&A
- ✓ Correctness in answer

- ✓ Multi-step reasoning
- ✓ Correctness in answer

ORQA

Benchmarking Results

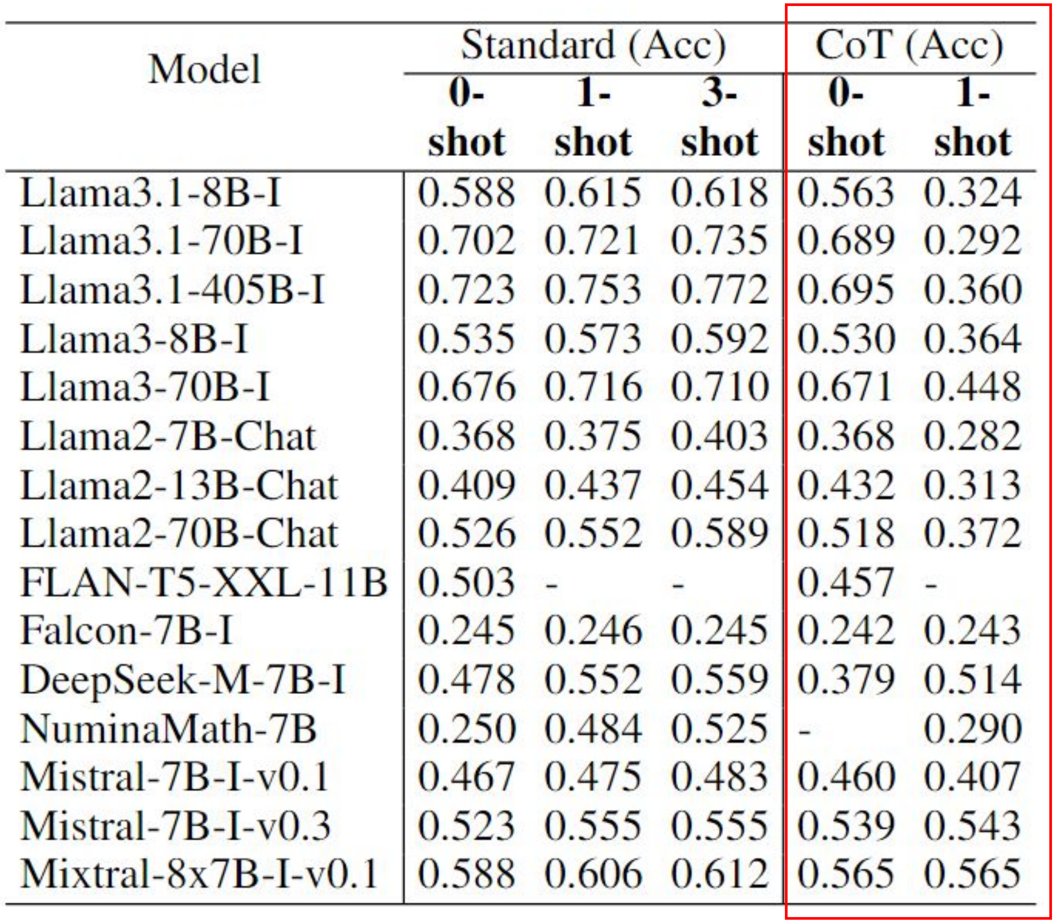


Model	Standard (Acc)			CoT (Acc)	
	0-shot	**1-shot**	**3-shot**	**0-shot**	**1-shot**
Llama3.1-8B-I	0.588	0.615	0.618	0.563	0.324
Llama3.1-70B-I	0.702	0.721	0.735	0.689	0.292
Llama3.1-405B-I	0.723	0.753	0.772	0.695	0.360
Llama3-8B-I	0.535	0.573	0.592	0.530	0.364
Llama3-70B-I	0.676	0.716	0.710	0.671	0.448
Llama2-7B-Chat	0.368	0.375	0.403	0.368	0.282
Llama2-13B-Chat	0.409	0.437	0.454	0.432	0.313
Llama2-70B-Chat	0.526	0.552	0.589	0.518	0.372
FLAN-T5-XXL-11B	0.503	-	-	0.457	-
Falcon-7B-I	0.245	0.246	0.245	0.242	0.243
DeepSeek-M-7B-I	0.478	0.552	0.559	0.379	0.514
NuminaMath-7B	0.250	0.484	0.525	-	0.290
Mistral-7B-I-v0.1	0.467	0.475	0.483	0.460	0.407
Mistral-7B-I-v0.3	0.523	0.555	0.555	0.539	0.543
Mixtral-8x7B-I-v0.1	0.588	0.606	0.612	0.565	0.565

0.93

Chain-of-Thought

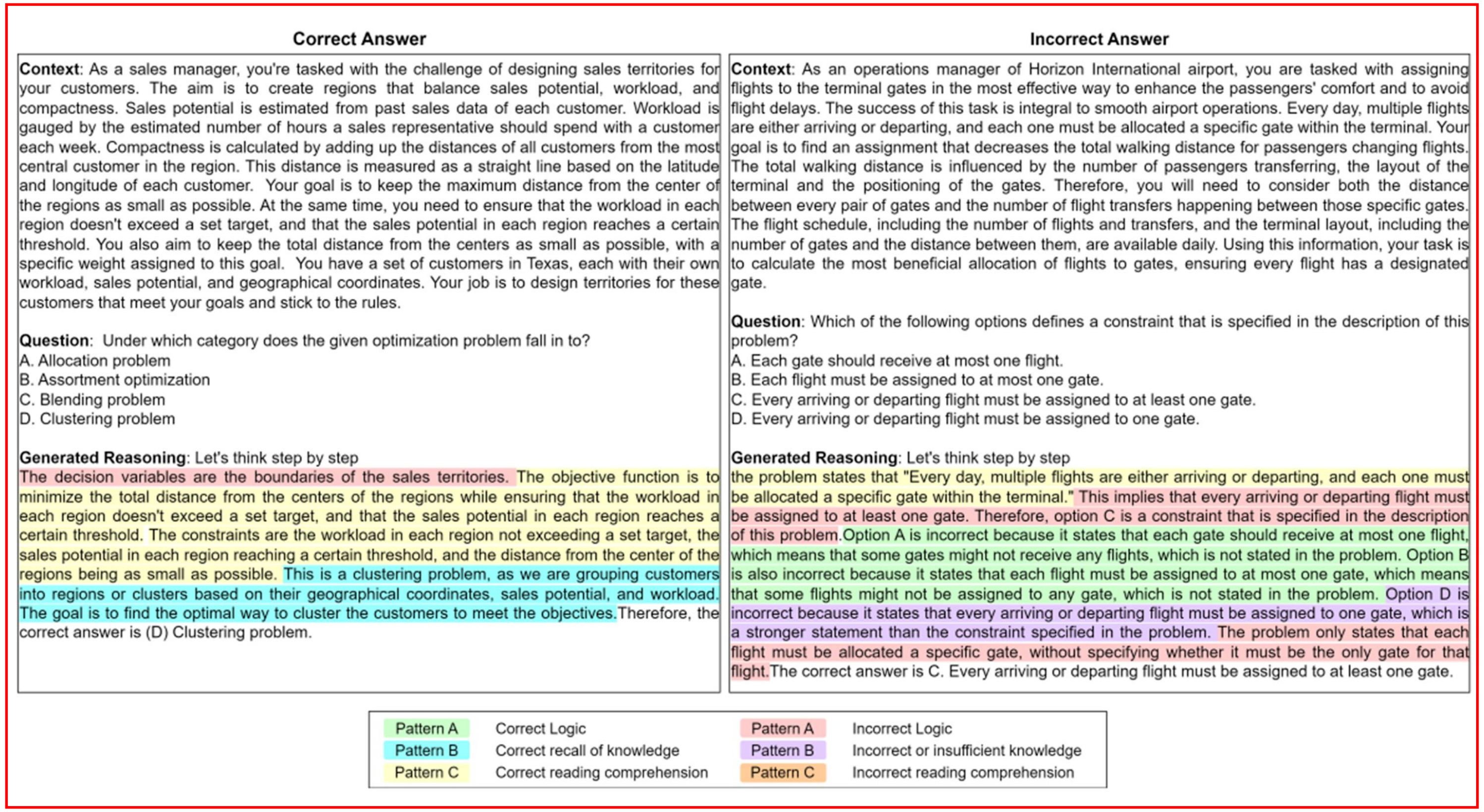
CoT had **negative** (or negligible) impact on model performance of most models



Model	Standard (Acc)			CoT (Acc)	
	0-shot	**1-shot**	**3-shot**	**0-shot**	**1-shot**
Llama3.1-8B-I	0.588	0.615	0.618	0.563	0.324
Llama3.1-70B-I	0.702	0.721	0.735	0.689	0.292
Llama3.1-405B-I	0.723	0.753	0.772	0.695	0.360
Llama3-8B-I	0.535	0.573	0.592	0.530	0.364
Llama3-70B-I	0.676	0.716	0.710	0.671	0.448
Llama2-7B-Chat	0.368	0.375	0.403	0.368	0.282
Llama2-13B-Chat	0.409	0.437	0.454	0.432	0.313
Llama2-70B-Chat	0.526	0.552	0.589	0.518	0.372
FLAN-T5-XXL-11B	0.503	-	-	0.457	-
Falcon-7B-I	0.245	0.246	0.245	0.242	0.243
DeepSeek-M-7B-I	0.478	0.552	0.559	0.379	0.514
NuminaMath-7B	0.250	0.484	0.525	-	0.290
Mistral-7B-I-v0.1	0.467	0.475	0.483	0.460	0.407
Mistral-7B-I-v0.3	0.523	0.555	0.555	0.539	0.543
Mixtral-8x7B-I-v0.1	0.588	0.606	0.612	0.565	0.565

Detailed Reasoning Analysis

Reasoning Analyses



Correct Answer

Context: As a sales manager, you're tasked with the challenge of designing sales territories for your customers. The aim is to create regions that balance sales potential, workload, and compactness. Sales potential is estimated from past sales data of each customer. Workload is gauged by the estimated number of hours a sales representative should spend with a customer each week. Compactness is calculated by adding up the distances of all customers from the most central customer in the region. This distance is measured as a straight line based on the latitude and longitude of each customer. Your goal is to keep the maximum distance from the center of the regions as small as possible. At the same time, you need to ensure that the workload in each region doesn't exceed a set target, and that the sales potential in each region reaches a certain threshold. You also aim to keep the total distance from the centers as small as possible, with a specific weight assigned to this goal. You have a set of customers in Texas, each with their own workload, sales potential, and geographical coordinates. Your job is to design territories for these customers that meet your goals and stick to the rules.

Question: Under which category does the given optimization problem fall in to?
A. Allocation problem
B. Assortment optimization
C. Blending problem
D. Clustering problem

Generated Reasoning: Let's think step by step
The decision variables are the boundaries of the sales territories. The objective function is to minimize the total distance from the centers of the regions while ensuring that the workload in each region doesn't exceed a set target, and that the sales potential in each region reaches a certain threshold. The constraints are the workload in each region not exceeding a set target, the sales potential in each region reaching a certain threshold, and the distance from the center of the regions being as small as possible. This is a clustering problem, as we are grouping customers into regions or clusters based on their geographical coordinates, sales potential, and workload. The goal is to find the optimal way to cluster the customers to meet the objectives.Therefore, the correct answer is (D) Clustering problem.

Incorrect Answer

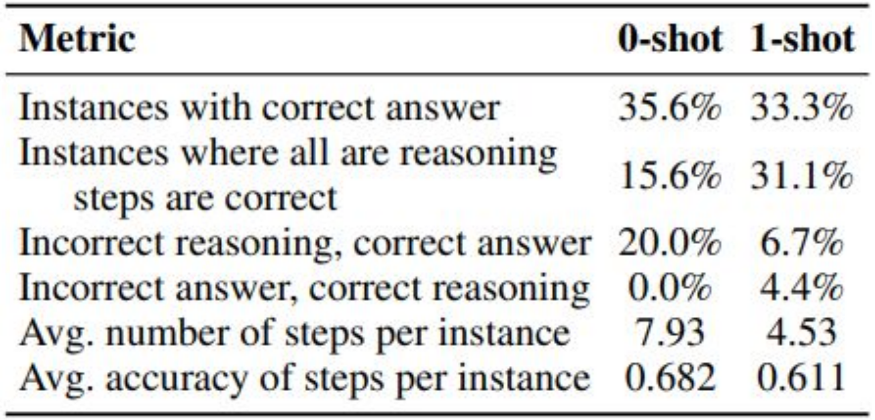
Context: As an operations manager of Horizon International airport, you are tasked with assigning flights to the terminal gates in the most effective way to enhance the passengers' comfort and to avoid flight delays. The success of this task is integral to smooth airport operations. Every day, multiple flights are either arriving or departing, and each one must be allocated a specific gate within the terminal. Your goal is to find an assignment that decreases the total walking distance for passengers changing flights. The total walking distance is influenced by the number of passengers transferring, the layout of the terminal and the positioning of the gates. Therefore, you will need to consider both the distance between every pair of gates and the number of flight transfers happening between those specific gates. The flight schedule, including the number of flights and transfers, and the terminal layout, including the number of gates and the distance between them, are available daily. Using this information, your task is to calculate the most beneficial allocation of flights to gates, ensuring every flight has a designated gate.

Question: Which of the following options defines a constraint that is specified in the description of this problem?
A. Each gate should receive at most one flight.
B. Each flight must be assigned to at most one gate.
C. Every arriving or departing flight must be assigned to at least one gate.
D. Every arriving or departing flight must be assigned to one gate.

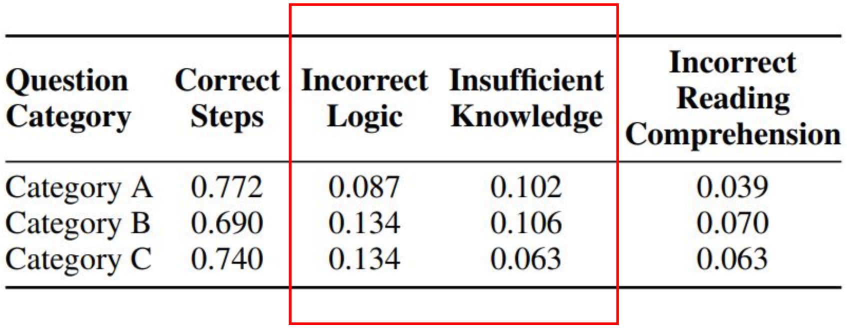
Generated Reasoning: Let's think step by step
the problem states that "Every day, multiple flights are either arriving or departing, and each one must be allocated a specific gate within the terminal." This implies that every arriving or departing flight must be assigned to at least one gate. Therefore, option C is a constraint that is specified in the description of this problem.Option A is incorrect because it states that each gate should receive at most one flight, which means that some gates might not receive any flights, which is not stated in the problem. Option B is also incorrect because it states that each flight must be assigned to at most one gate, which means that some flights might not be assigned to any gate, which is not stated in the problem. Option D is incorrect because it states that every arriving or departing flight must be assigned to one gate, which is a stronger statement than the constraint specified in the problem. The problem only states that each flight must be allocated a specific gate, without specifying whether it must be the only gate for that flight.The correct answer is C. Every arriving or departing flight must be assigned to at least one gate.

Pattern A	Correct Logic	Pattern A	Incorrect Logic
Pattern B	Correct recall of knowledge	Pattern B	Incorrect or insufficient knowledge
Pattern C	Correct reading comprehension	Pattern C	Incorrect reading comprehension

Reasoning Insights (stats)



Metric	0-shot	1-shot
Instances with correct answer	35.6%	33.3%
Instances where all are reasoning steps are correct	15.6%	31.1%
Incorrect reasoning, correct answer	20.0%	6.7%
Incorrect answer, correct reasoning	0.0%	4.4%
Avg. number of steps per instance	7.93	4.53
Avg. accuracy of steps per instance	0.682	0.611

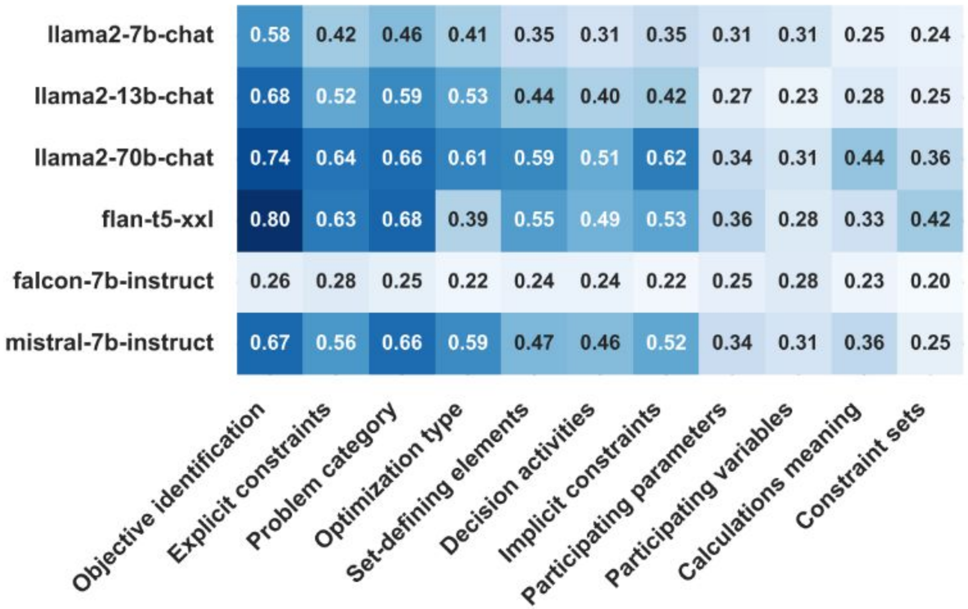


Question Category	Correct Steps	Incorrect Logic	Insufficient Knowledge	Incorrect Reading Comprehension
Category A	0.772	0.087	0.102	0.039
Category B	0.690	0.134	0.106	0.070
Category C	0.740	0.134	0.063	0.063

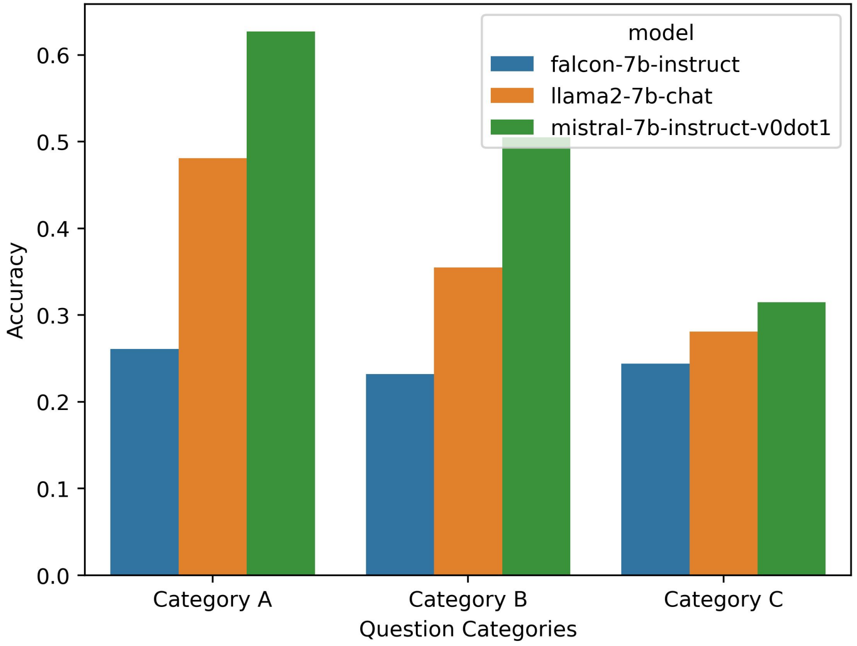
Question Types (Difficulty)

Significantly more difficult:

- **Model building & understanding** are significantly more difficult
- **Relationships** of optimization model components

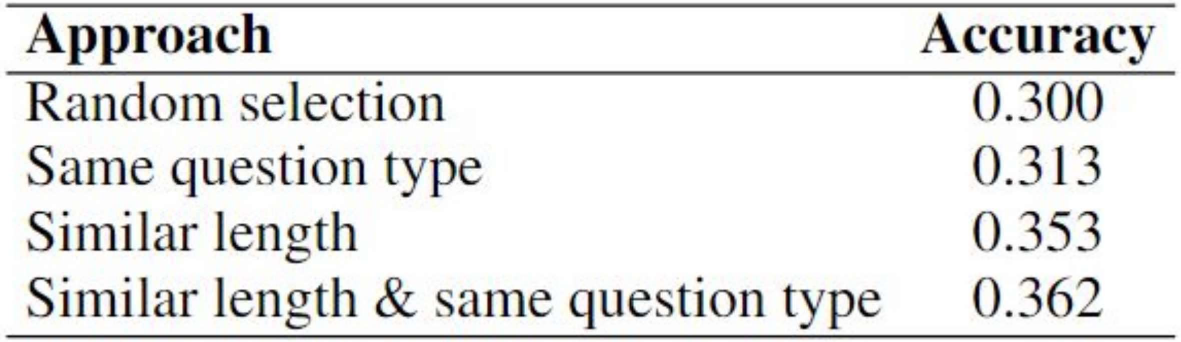


	Objective identification	Explicit constraints	Problem category	Optimization type	Set-defining elements	Decision activities	Implicit constraints	Participating parameters	Participating variables	Calculations meaning	Constraint sets
llama2-7b-chat	0.58	0.42	0.46	0.41	0.35	0.31	0.35	0.31	0.31	0.25	0.24
llama2-13b-chat	0.68	0.52	0.59	0.53	0.44	0.40	0.42	0.27	0.23	0.28	0.25
llama2-70b-chat	0.74	0.64	0.66	0.61	0.59	0.51	0.62	0.34	0.31	0.44	0.36
flan-t5-xxl	0.80	0.63	0.68	0.39	0.55	0.49	0.53	0.36	0.28	0.33	0.42
falcon-7b-instruct	0.26	0.28	0.25	0.22	0.24	0.24	0.22	0.25	0.28	0.23	0.20
mistral-7b-instruct	0.67	0.56	0.66	0.59	0.47	0.46	0.52	0.34	0.31	0.36	0.25



ICL and Trigger Prompt on Reasoning

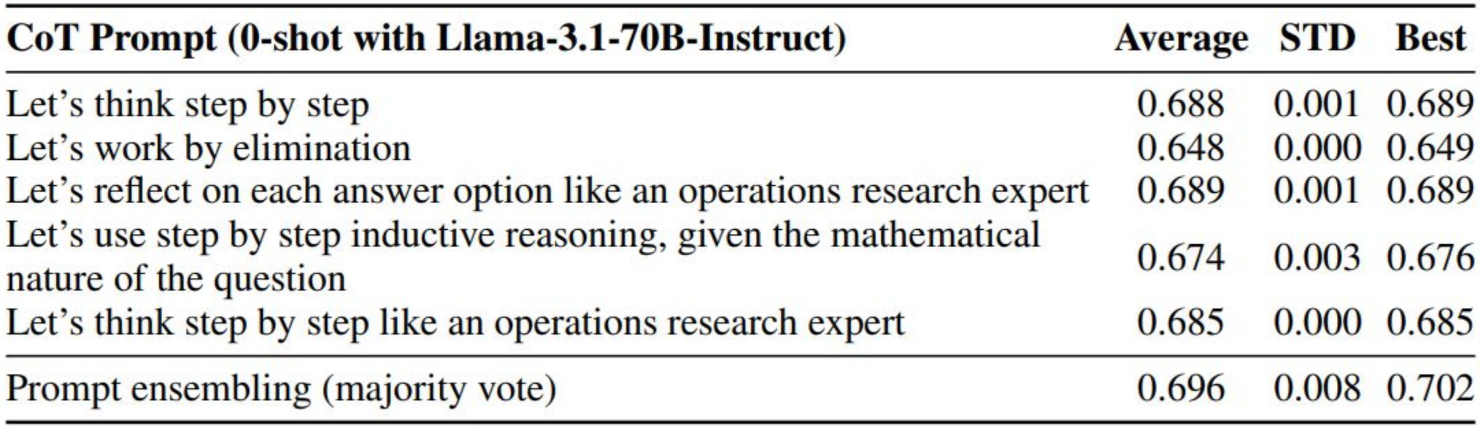
ICL Selection



Approach	Accuracy
Random selection	0.300
Same question type	0.313
Similar length	0.353
Similar length & same question type	0.362

Length of ICL examples has more impact on performance than question type (similarity)

Trigger Prompts Selection

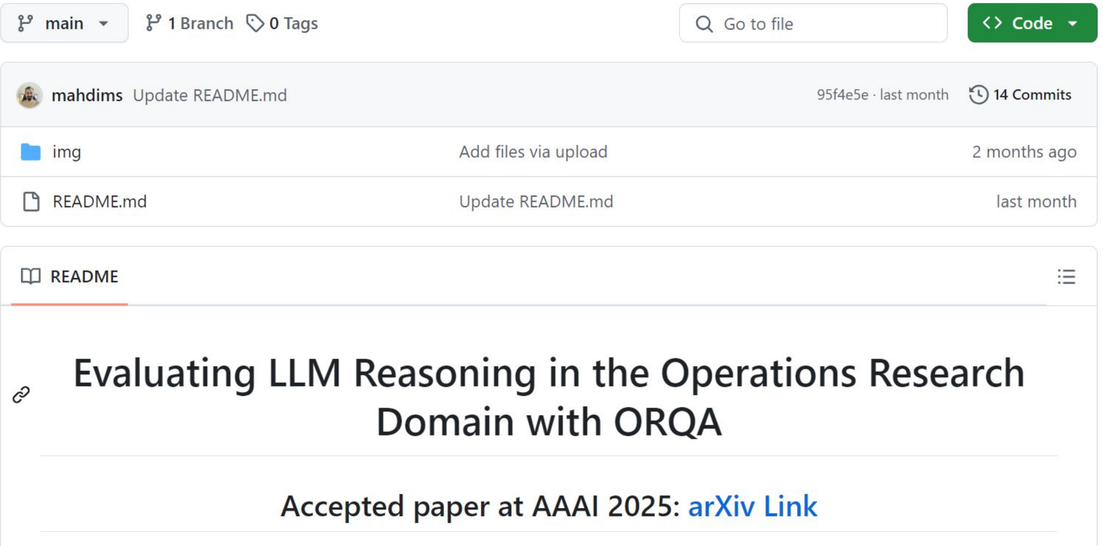


CoT Prompt (0-shot with Llama-3.1-70B-Instruct)	Average	STD	Best
Let’s think step by step	0.688	0.001	0.689
Let’s work by elimination	0.648	0.000	0.649
Let’s reflect on each answer option like an operations research expert	0.689	0.001	0.689
Let’s use step by step inductive reasoning, given the mathematical nature of the question	0.674	0.003	0.676
Let’s think step by step like an operations research expert	0.685	0.000	0.685
Prompt ensembling (majority vote)	0.696	0.008	0.702

Trigger prompts and ensembling multiple runs may yield better results

Conclusion & Future Works

https://github.com/nl4opt/ORQA



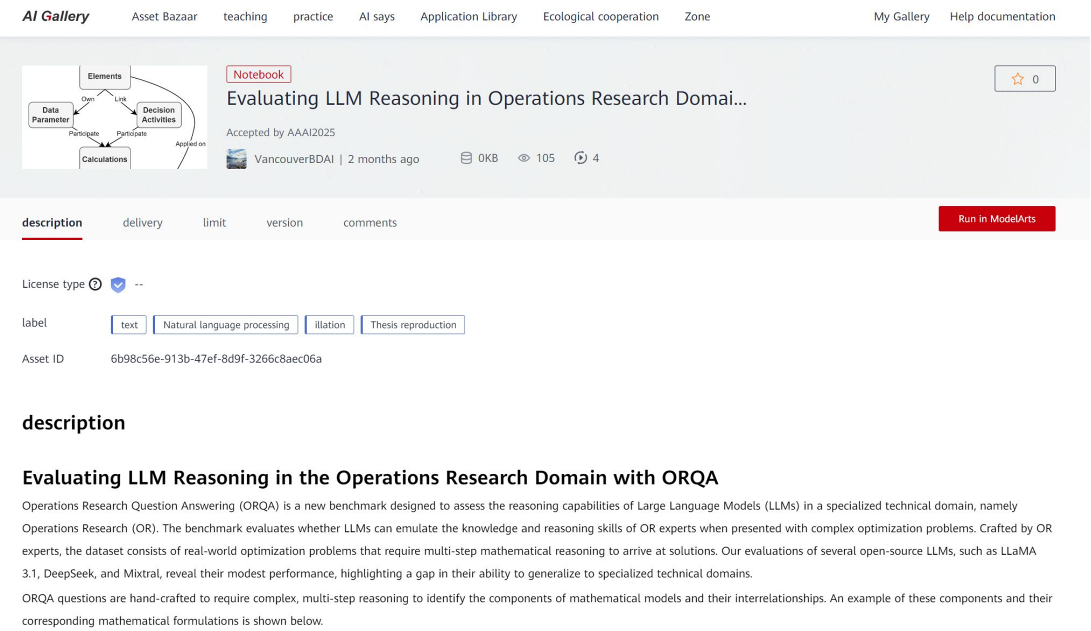
mahdims Update README.md 95f4e5e · last month 14 Commits

img	Add files via upload	2 months ago
README.md	Update README.md	last month

Evaluating LLM Reasoning in the Operations Research Domain with ORQA

Accepted paper at AAAI 2025: arXiv Link

AI Gallery_Notebook详情_开发者_华为云 (huaweicloud.com)



Evaluating LLM Reasoning in Operations Research Domai...

Accepted by AAAI2025

VancouverBDAI | 2 months ago 0KB 105 4

License type --

label text Natural language processing illation Thesis reproduction

Asset ID 6b98c56e-913b-47ef-8d9f-3266c8aec06a

description

Evaluating LLM Reasoning in the Operations Research Domain with ORQA

Operations Research Question Answering (ORQA) is a new benchmark designed to assess the reasoning capabilities of Large Language Models (LLMs) in a specialized technical domain, namely Operations Research (OR). The benchmark evaluates whether LLMs can emulate the knowledge and reasoning skills of OR experts when presented with complex optimization problems. Crafted by OR experts, the dataset consists of real-world optimization problems that require multi-step mathematical reasoning to arrive at solutions. Our evaluations of several open-source LLMs, such as LLaMA 3.1, DeepSeek, and Mixtral, reveal their modest performance, highlighting a gap in their ability to generalize to specialized technical domains.

ORQA questions are hand-crafted to require complex, multi-step reasoning to identify the components of mathematical models and their interrelationships. An example of these components and their corresponding mathematical formulations is shown below.

Future Works

- Evaluate on more models
- Expand the dataset
- Automate the reasoning analysis

ORQA □ Models more proficient at reasoning

deepseek

QwQ-32B-Preview

o3-mini

Thank you!

Evaluating LLM Reasoning in the Operations Research Domain with ORQA

Mahdi Mostajabdaveh, Timothy Tin Long Yu, Samarendra Chandan Bindu Dash, Rindra Ramamonjison,

Jabo Serge Byusa, Giuseppe Carenini, Zirui Zhou, Yong Zhang