Institute for Enterprise Systems
Empowering Digital Transformation

UNIVERSITY OF MANNHEIM

European MTP – Intelligent Hospital Logistics

Agenda

Project Goal Recap

Progress Update

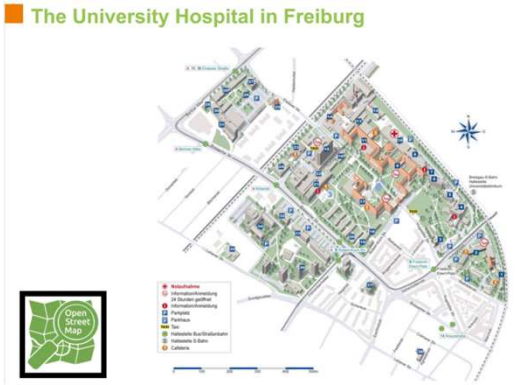
Challenges and Solutions

Outlook

Project Goal Recap

Transport logistics platform for autonomous vehicles on clinic premises

Simulation environment – University Hospital Freiburg



Research Goal: **Handle Incidents**

Can LLMs **simulate** the **impact** of **incidents** on routing graphs?

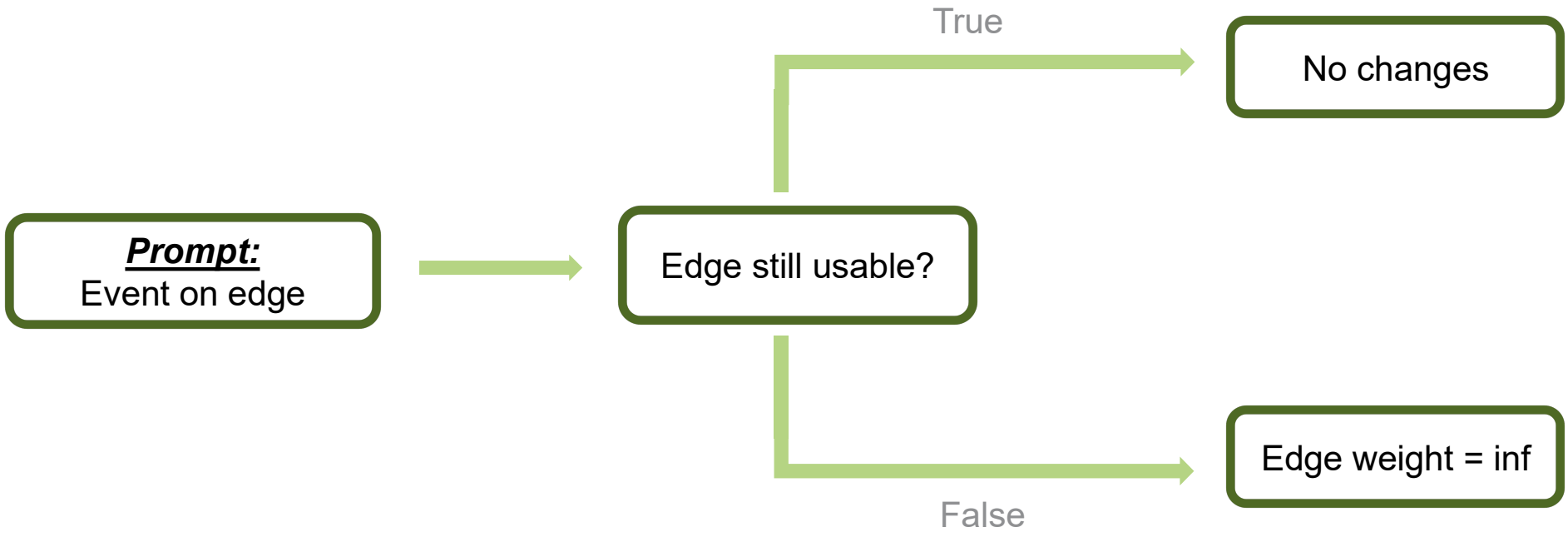
Progress Update

Previous Status: Independent Submodels

Approach 1: Edge Usability

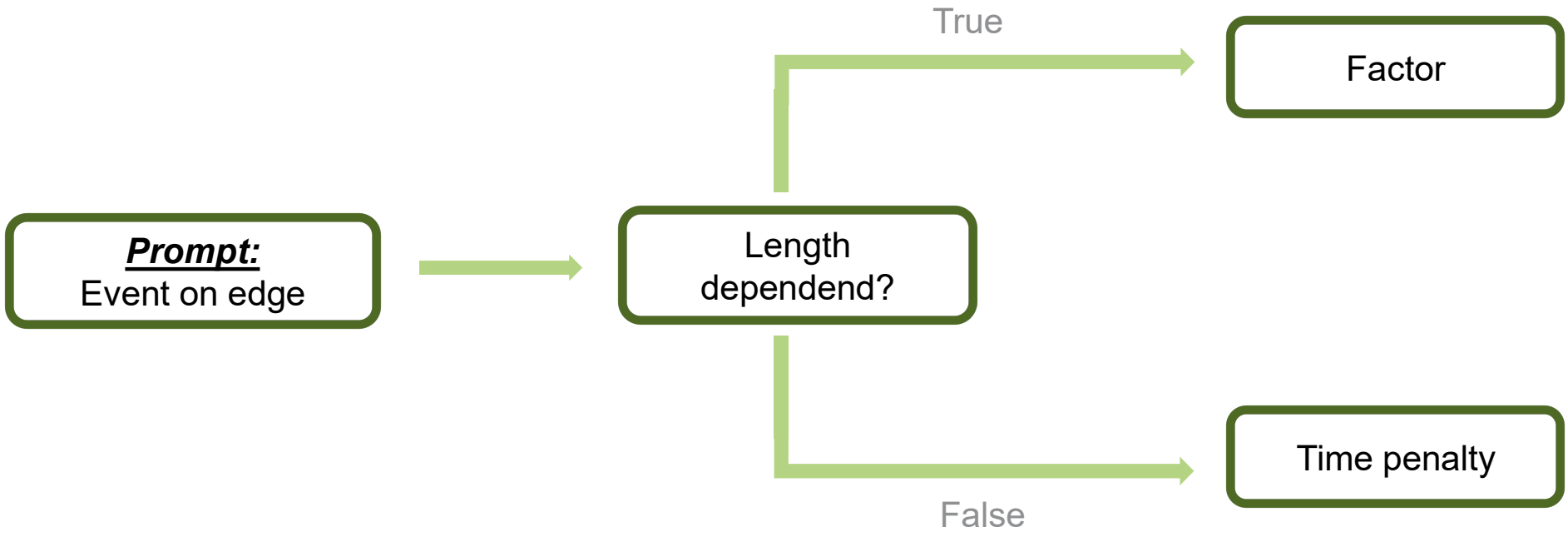
Zero- and Few-Shot

Zero- and Few-Shot



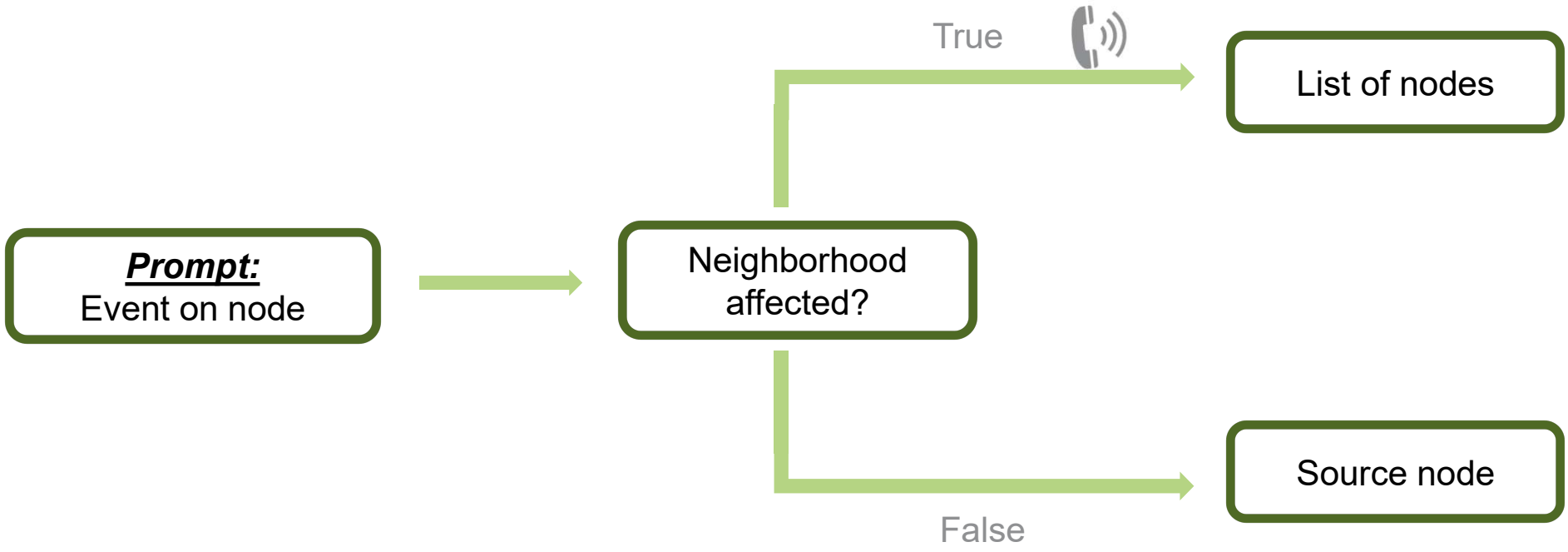
Previous Status: Independent Submodels

Approach 2: Dynamic Edge Weights



Previous Status: Independent Submodels

Approach 3: Function Calling



We built a meta-model to increase usability and accuracy after intermediate presentation

Goals after intermediate presentation

Increase usability
No need to choose model

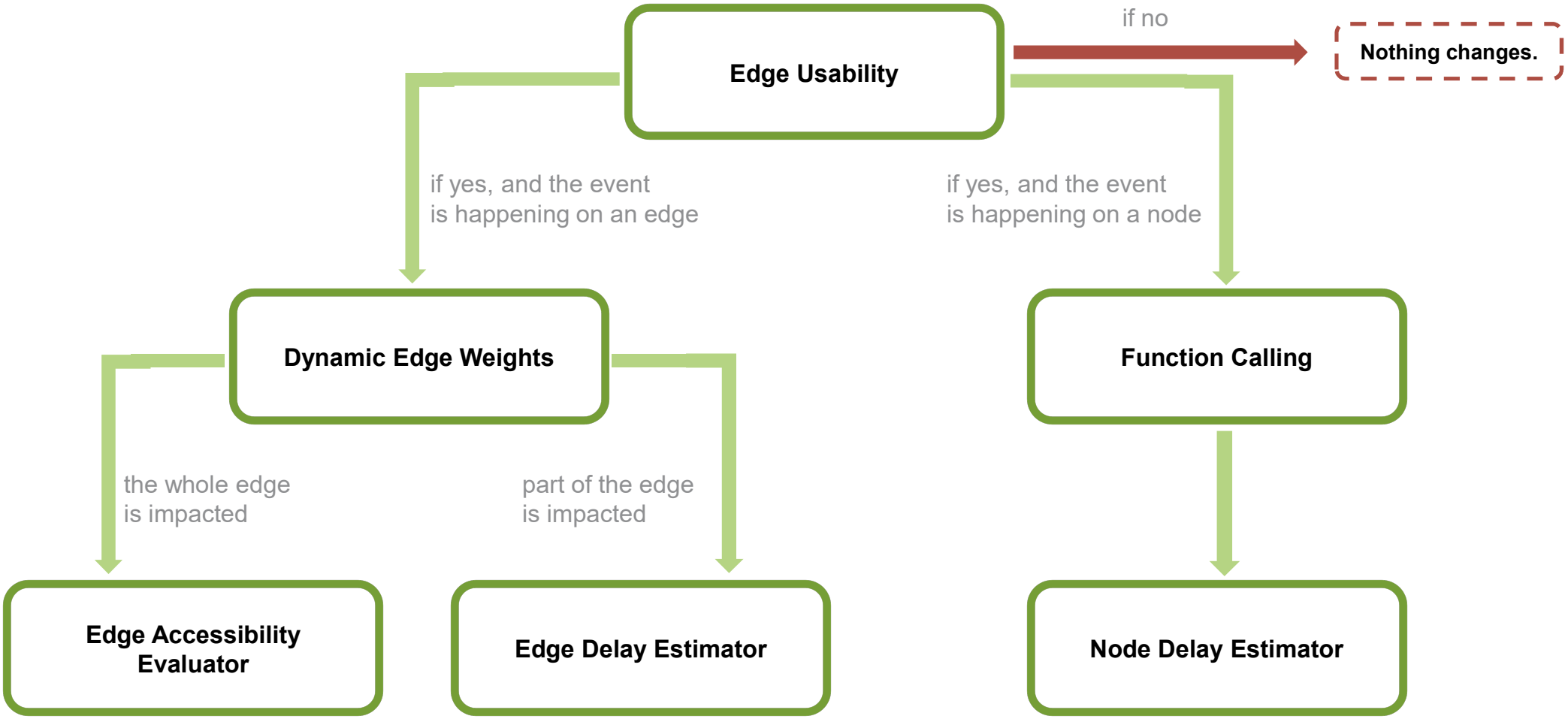
Increase awareness of environment
Also consider neighborhood of incident

Increase detection accuracy
Incident detection more accurate

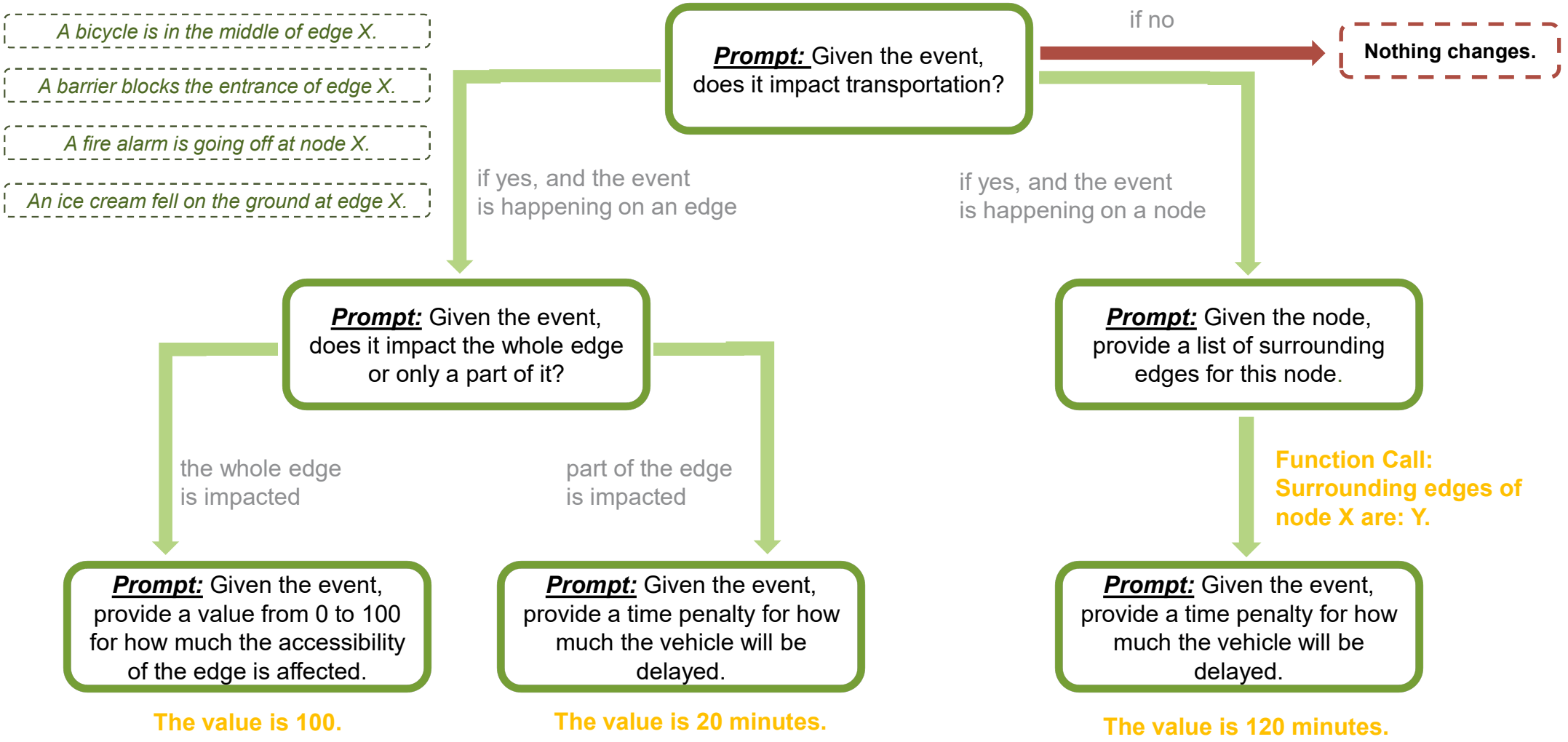
Progress

- Build **meta-model** that combines best approaches
- Introduce **time penalty** for congestion on diversion
- Increase **precision** of incident detection

Progress Update: Meta-Model



Progress Update: Meta-Model



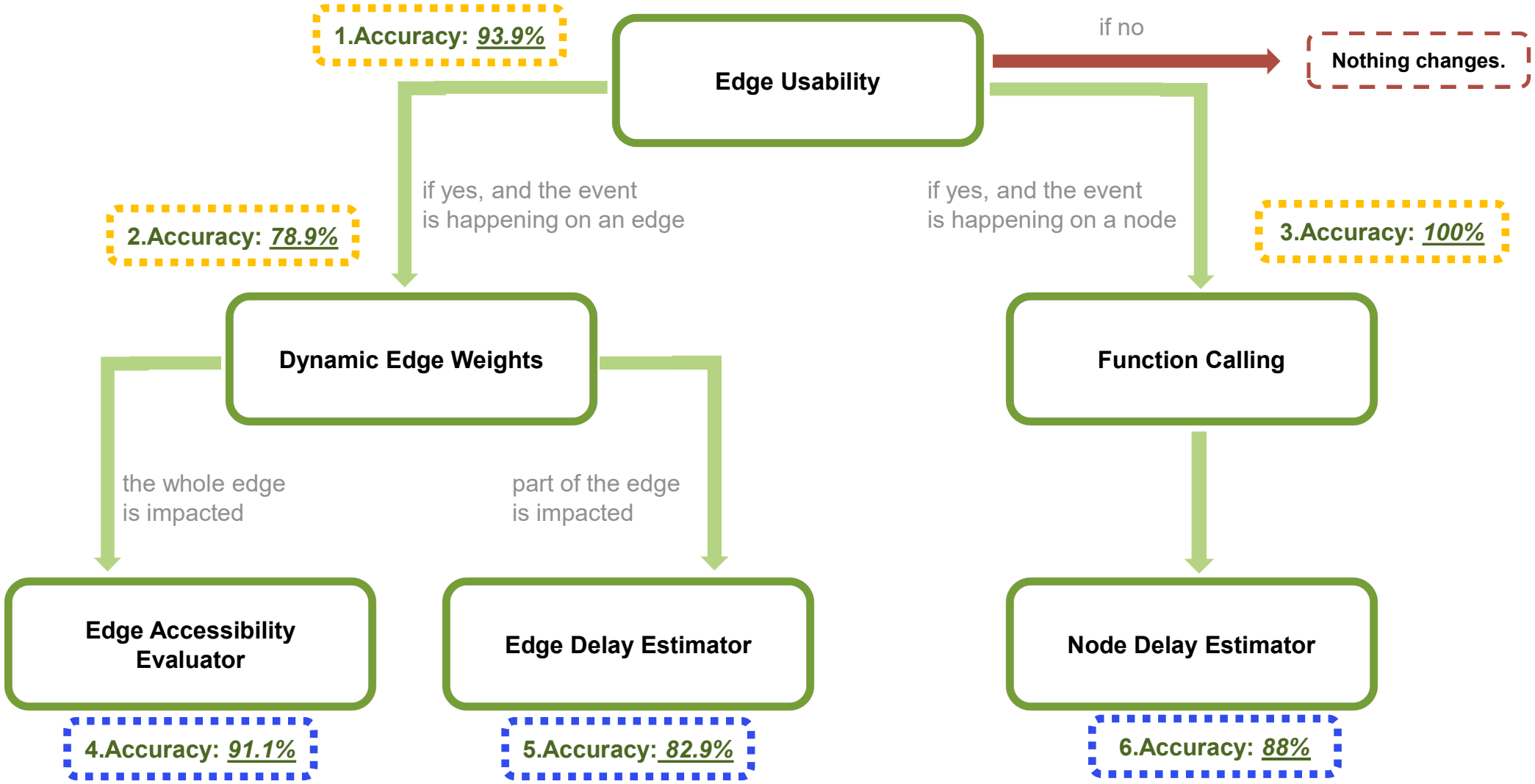
Evaluation of our Meta-Model

Reference-based Evaluation

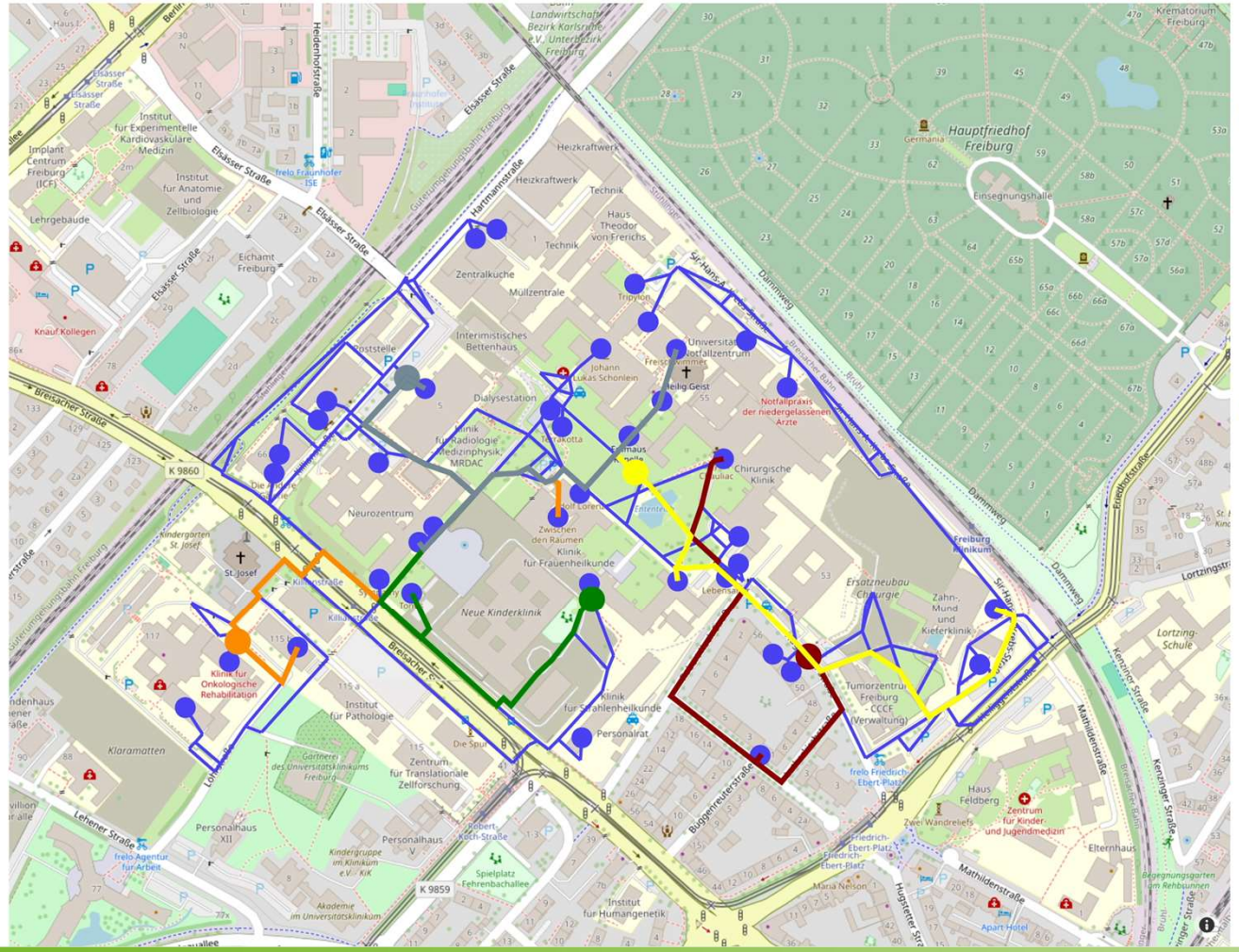
- Applied when a ground truth is available.
- The output of the LLMs is compared to the ground truth.
- Used for our Edge Usability, Dynamic Edge Weights & Function Calling models.

Reference-free Evaluation

- Applied when the ground truth is not available.
- 3 different Evaluation LLMs are employed and tasked to evaluate the responses of our previous LLMs.
- The event and factor/time penalty given as input to the Evaluation LLMs to evaluate how accurate the factor/time penalty is.
- Used for our Edge Accessibility Evaluator, Edge Delay Estimator & Node Delay Estimator models.


1.Accuracy: 93.9%
Edge Usability
if no
Nothing changes.
if yes, and the event is happening on an edge
if yes, and the event is happening on a node
2.Accuracy: 78.9%
3.Accuracy: 100%
Dynamic Edge Weights
Function Calling
the whole edge is impacted
part of the edge is impacted
Edge Accessibility Evaluator
Edge Delay Estimator
Node Delay Estimator
4.Accuracy: 91.1%
5.Accuracy: 82.9%
6.Accuracy: 88%

Intelligent Hospital Logistics



Incidents | Events | Status | Prompts

Orders

source	target	items	status	
Klinik für Frauenheilkunde	Gum	Masks	in progress...	
Christian Daniel Nussbaum	Theodor Billroth	Water	in progress...	
Freischwimmer	Neurozentrum	Palenka	in progress...	lig
Blutspende Freiburg	Zwischen den Räumen	Blood	in progress...	
Augenklinik / HNO	Tonus	Blood	in progress...	

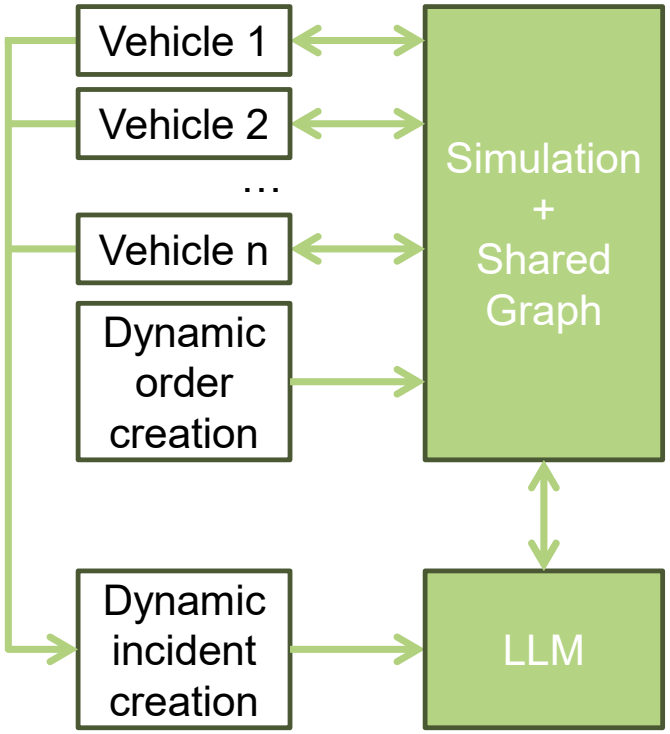
Vehicles

position	status	targetNode	Vehicle
(48.00490760, 7.83332430)	moving	7112443020	darkorange
(48.00525588, 7.83760610)	moving	2401035988	green
(48.00631622, 7.83824530)	moving	7126868392	yellow
(48.00477937, 7.84038710)	moving	9622545345	darkred
(48.00709089, 7.83531800)	moving	1280377565	lightslategray

Challenges and Solutions

Challenge 1 – Our distributed system is a challenge for data consistency

Challenge



- Modules have independent runtime
- Creation of incident & orders is dynamic
- Communication and LLM introduce latency
- Movement of vehicles happens in parallel

Decoupled system is a challenge for data consistency

Solution

- Use MQTT as message protocol to guarantee message delivery
- Parallelize routing, creation of incidents and LLM
- Establish graph as a shared single source of truth
- Reduce data duplication to absolute minimum

Parallel computing and minimal data duplication ensures consistency

Challenge 2 – Handling Large Language Models

Challenge

Solution

Highly dynamic topic with fast progress in research and very little documentation.	→	Continuously explore new features and models, using trial and error to optimize LLMs for handling incidents in a graph structure.
Restricted access to LLMs due to API-based requirements and associated financial costs.	→	Utilize LLMs that are easy to access with an API key and have a favorable cost-performance ratio.
Deploying LLMs locally requires significant computational resources and can result in increased latency.	→	Using university servers to run LLMs substantially reduces latency and improves overall performance.

Challenge 2 – Handling Large Language Models

Challenge

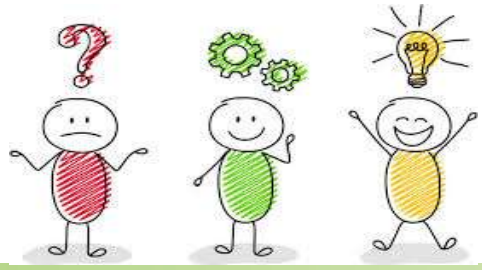
LLMs are not trained to understand graph structures and it remains an active research area.

Difficult evaluating LLM responses in a numerical form where no ground truth is available.

Solution

Leverage function calling to enable LLMs to effectively interact with and query graph data.

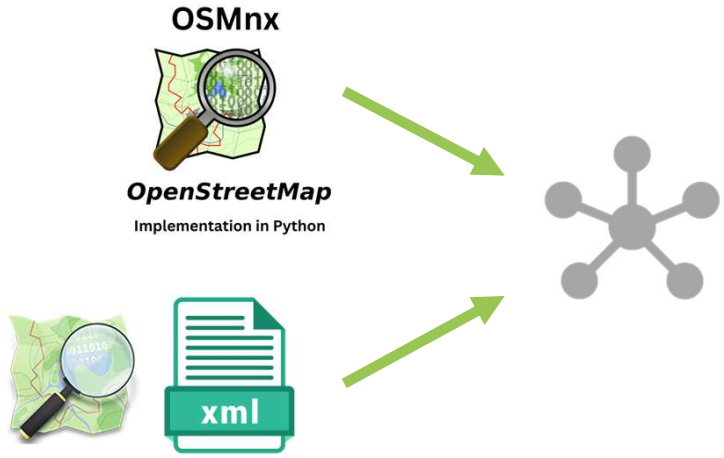
Implement a reference-free evaluation approach by using a secondary LLM to assess the quality of responses from primary LLMs.



Challenge 3 – Graph Construction

Challenge 3.1: Usage of two different sources with different edge definitions.

Solution: Use OSMnx graph with fewer edges and search in ways of nx graph for possible matches.



Challenge 3.2: Boundaries for map extraction from OSM are rectangles.

Solution: Start with larger map and cleverly drop some parts of the graph.

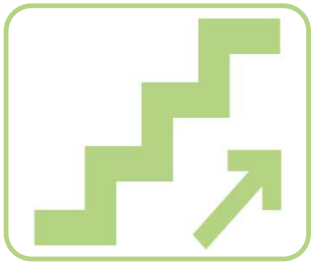
Outlook

1) Prioritize Goods

- Can LLMs be used for solving **routing** problems across **heterogeneous goods?**

2) Combine Orders

- Can orders be **joined** together to **save milage** by serving multiple destinations with **one ride**?



3) Transportation inside Buildings

- Can the simulation also include **transportation inside buildings** and account for **different** types of **obstacles** such as elevators, stairs etc.?

Thank You For Your Attention !

Do You Have Any Questions ?