Prompting or Fine-tuning? A Comparative Study of Large Language Models for Taxonomy Construction

Boqi Chen[1], Fandi Yi[1], Daniel Varro[1,2]

[1]Department of Electrical & Computer Engineering, McGill University
[2]Department of Computer and Information Science, Linköping University

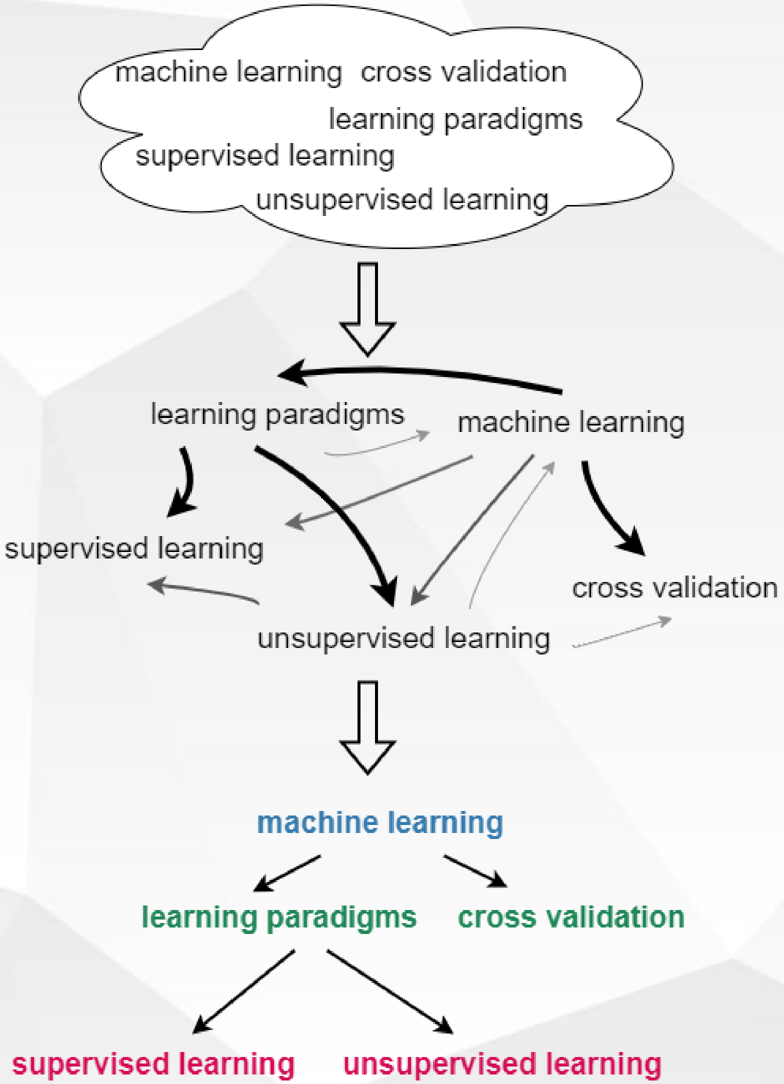
boqi.chen@mail.mcgill.ca, fandi.yi@mail.mcgill.ca, daniel.varro@liu.se

Background : Taxonomy

- **Taxonomies** represent hierarchical relations between concepts or entities.

- Taxonomies are important in software engineering
  - **domain modeling**.
  - **object-oriented languages.**
  - **semantic web applications.**

- **Taxonomy construction** is identifying the hierarchical relations between set of concepts
  - **parent-child:** generalization
  - **inclusion relations:** composition

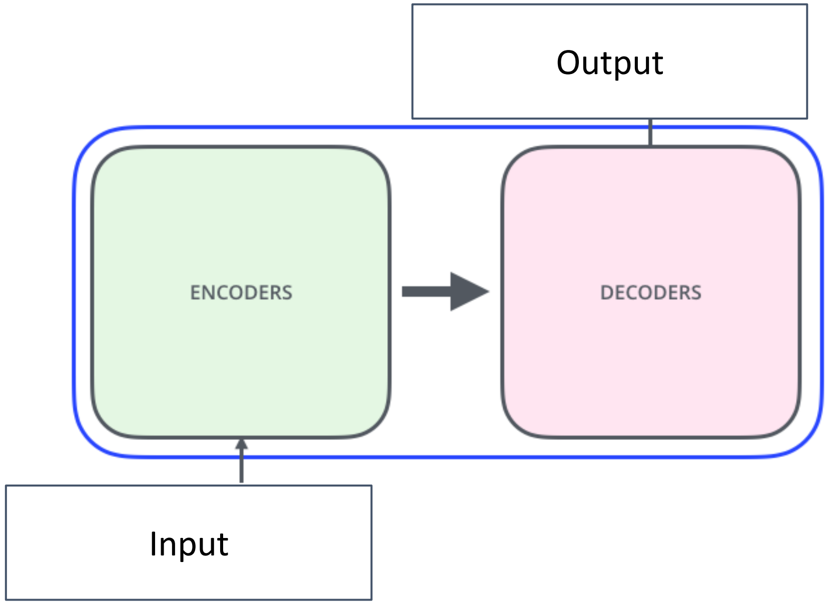


Background: Large Language Model (LLMs)

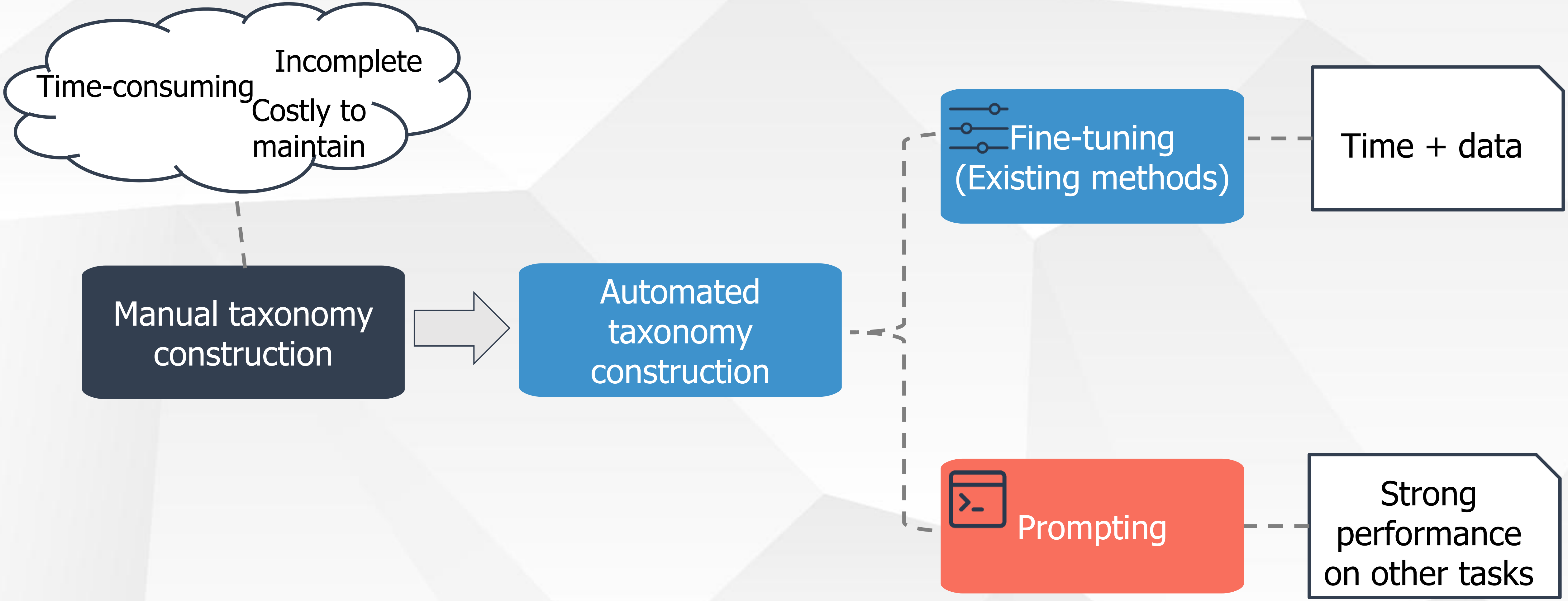
- **Large language models (LLMs)** are natural language processing methods for **text generation**

- For a sequence of input tokens (prompt), LLMs estimate the **probability** of the **next token**

- There are **two** methods for using pre-trained LLMs:
  - **Fine-tuning**: adapt with a task specific dataset
  - **Prompting:** provide instructions and examples as input for the task



Motivation: Explore LLM for Taxonomy



Time-consuming

Incomplete

Costly to maintain

Manual taxonomy construction

Automated taxonomy construction

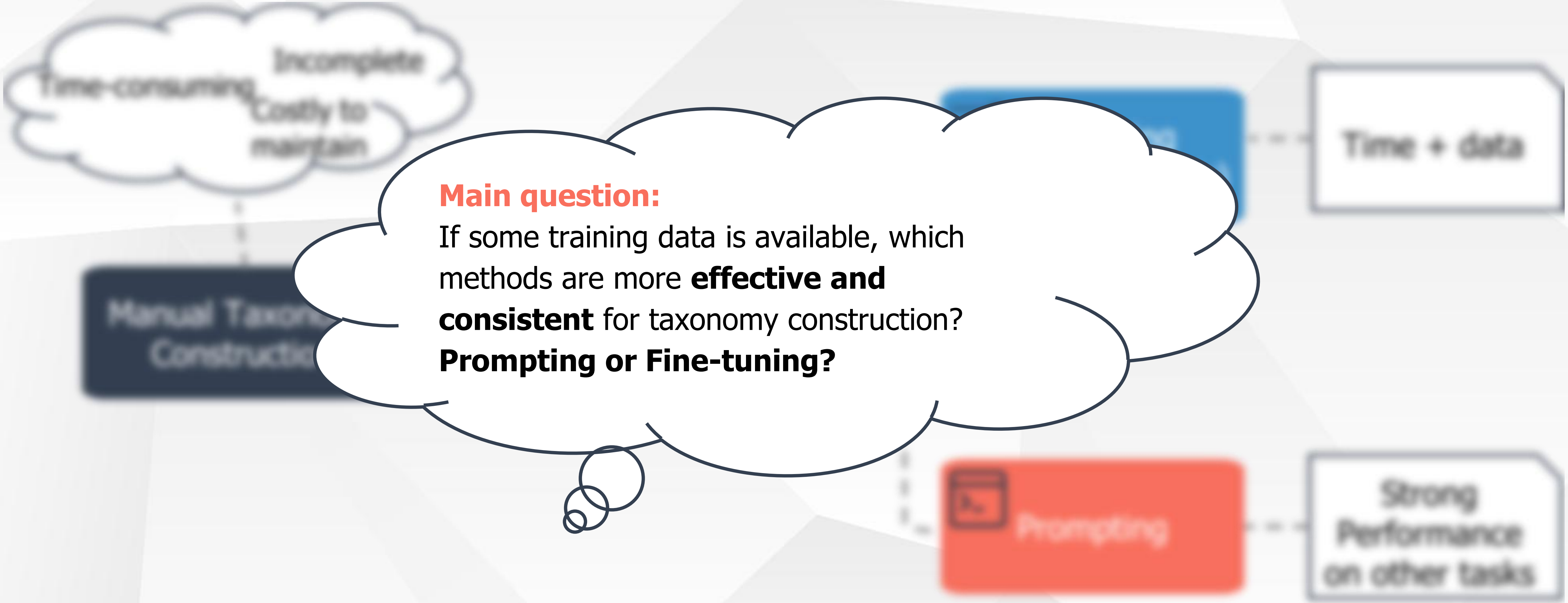
Fine-tuning (Existing methods)

Time + data

Prompting

Strong performance on other tasks

Motivation: Explore LLM for Taxonomy



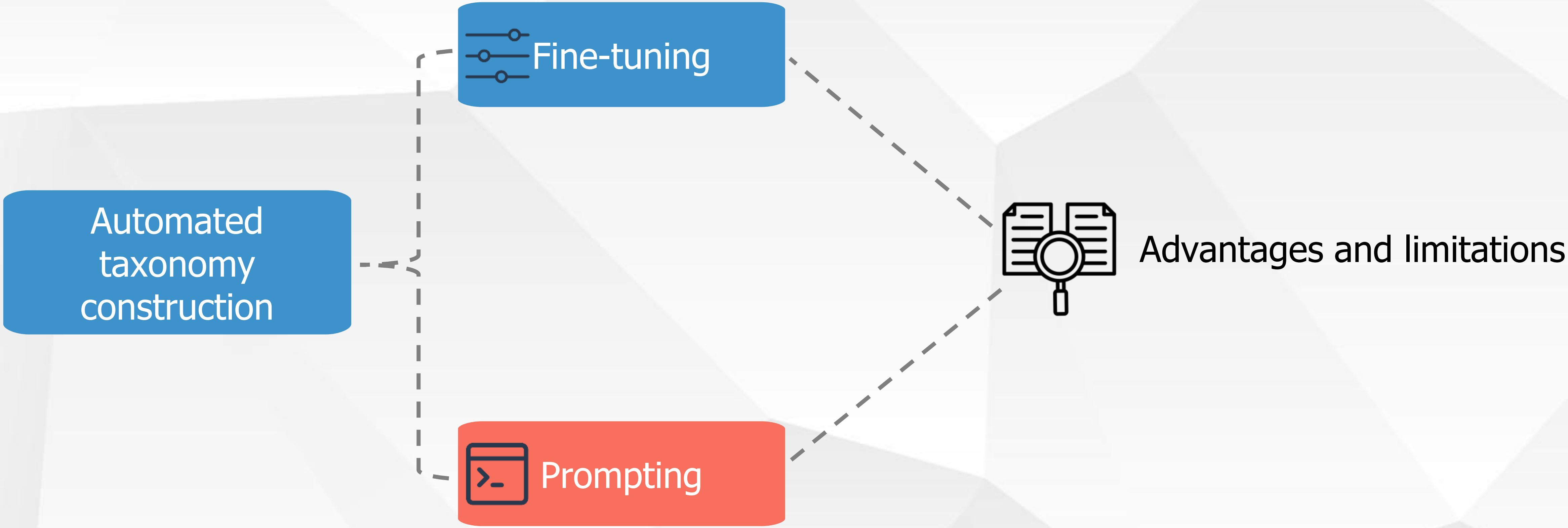
Main question:

If some training data is available, which methods are more **effective and consistent** for taxonomy construction?

Prompting or Fine-tuning?

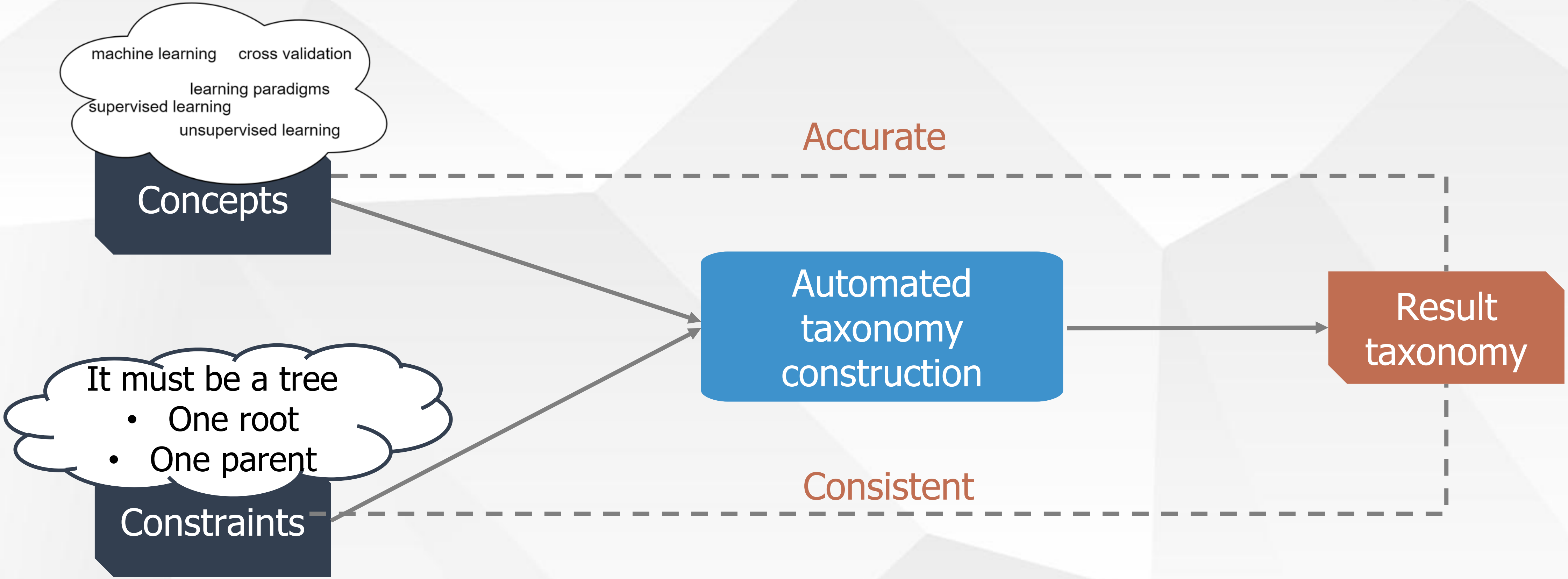
Objective

We present a **comparative study** using **LLMs** for taxonomy construction

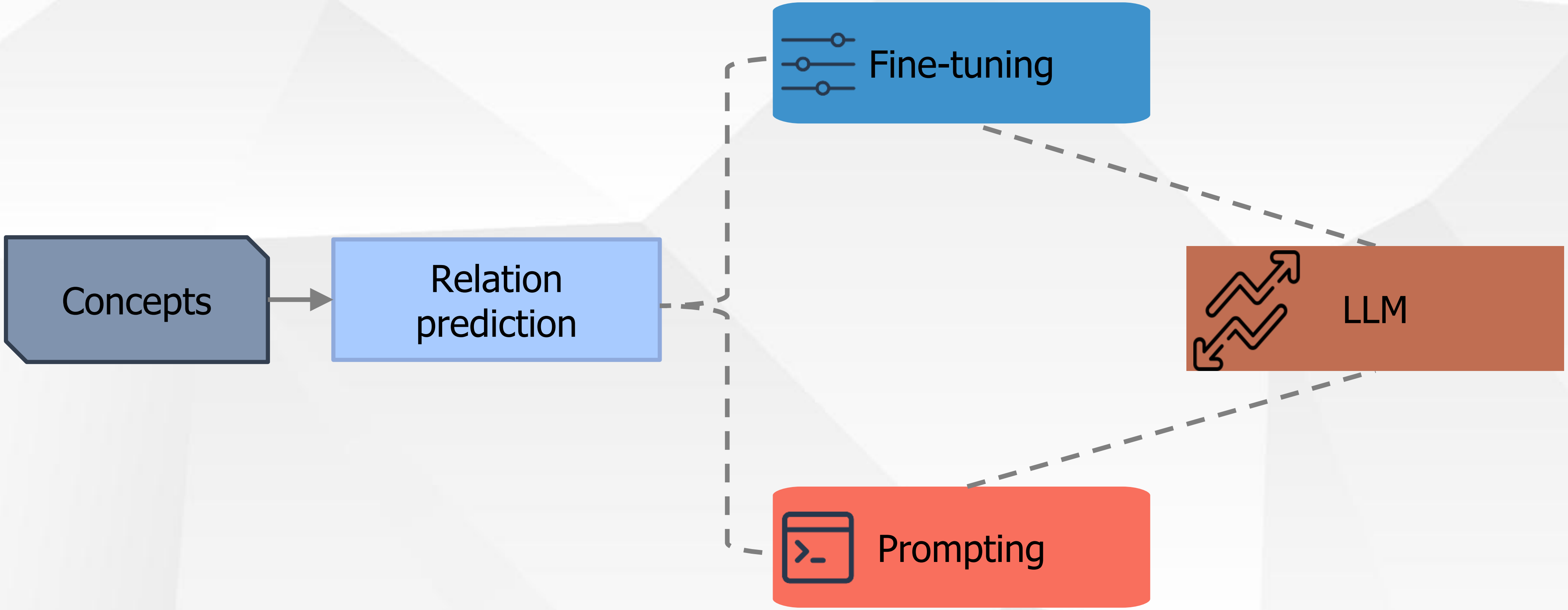


Problem Formulation

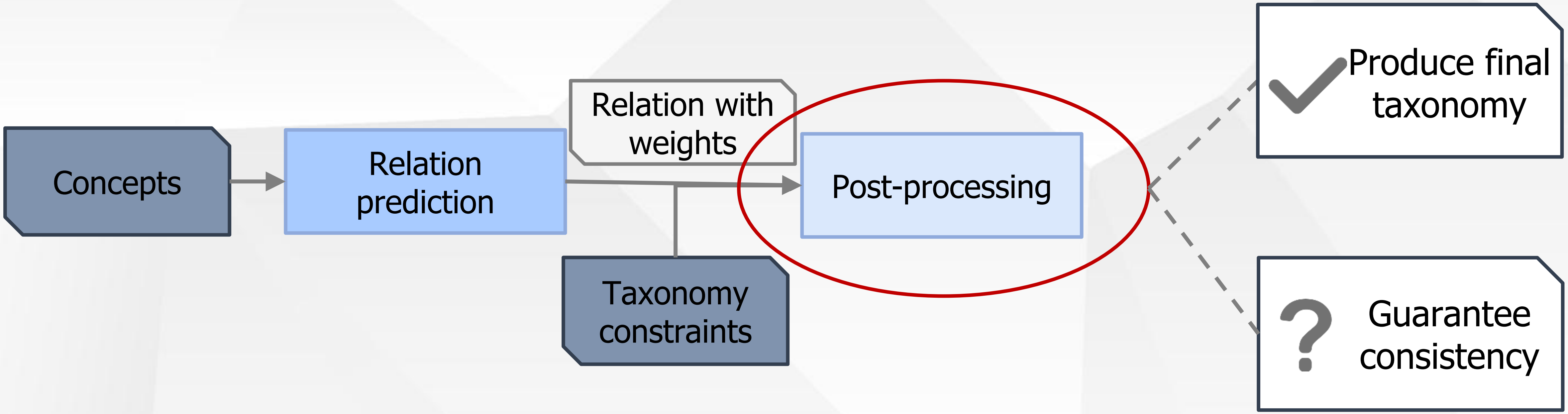
Given a set of concepts and constraint, create a taxonomy follows the constraints



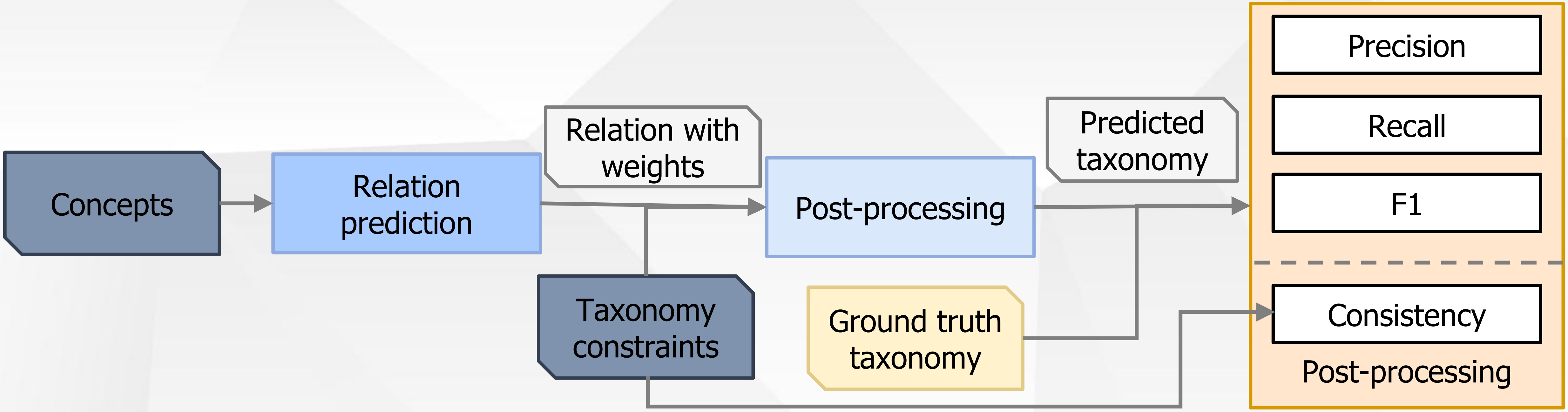
Approach Overview



Approach Overview

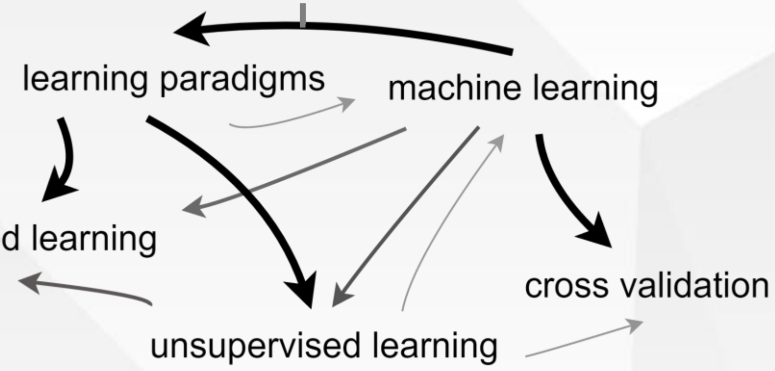


Approach Overview



Relation Prediction

Relation prediction



Edge weights: the likelihood of concept A being a parent of concept B

Candidate relation with weights

Relation Prediction

LINKÖPING UNIVERSITY

Fine-tuning → Predict weight for all relations

Method 1: layer-wise

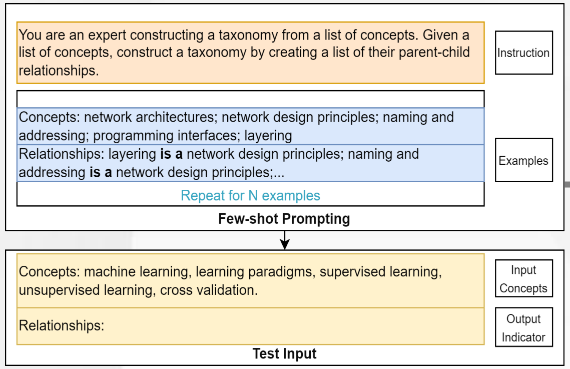
Train a subset of the parameters by selecting a few layers in the LLM

LLM

Method 2: LoRA

Update all parameters with low-rank adaptation (Reduces # parameters during training)

Prompting → GPT3.5 can be costly to run for all relations → Predict candidate relations directly



You are an expert constructing a taxonomy from a list of concepts. Given a list of concepts, construct a taxonomy by creating a list of their parent-child relationships.

Concepts: network architectures; network design principles; naming and addressing; programming interfaces; layering

Relationships: layering **is a** network design principles; naming and addressing **is a** network design principles;...

Repeat for N examples

Few-shot Prompting

Concepts: machine learning, learning paradigms, supervised learning, unsupervised learning, cross validation.

Relationships:

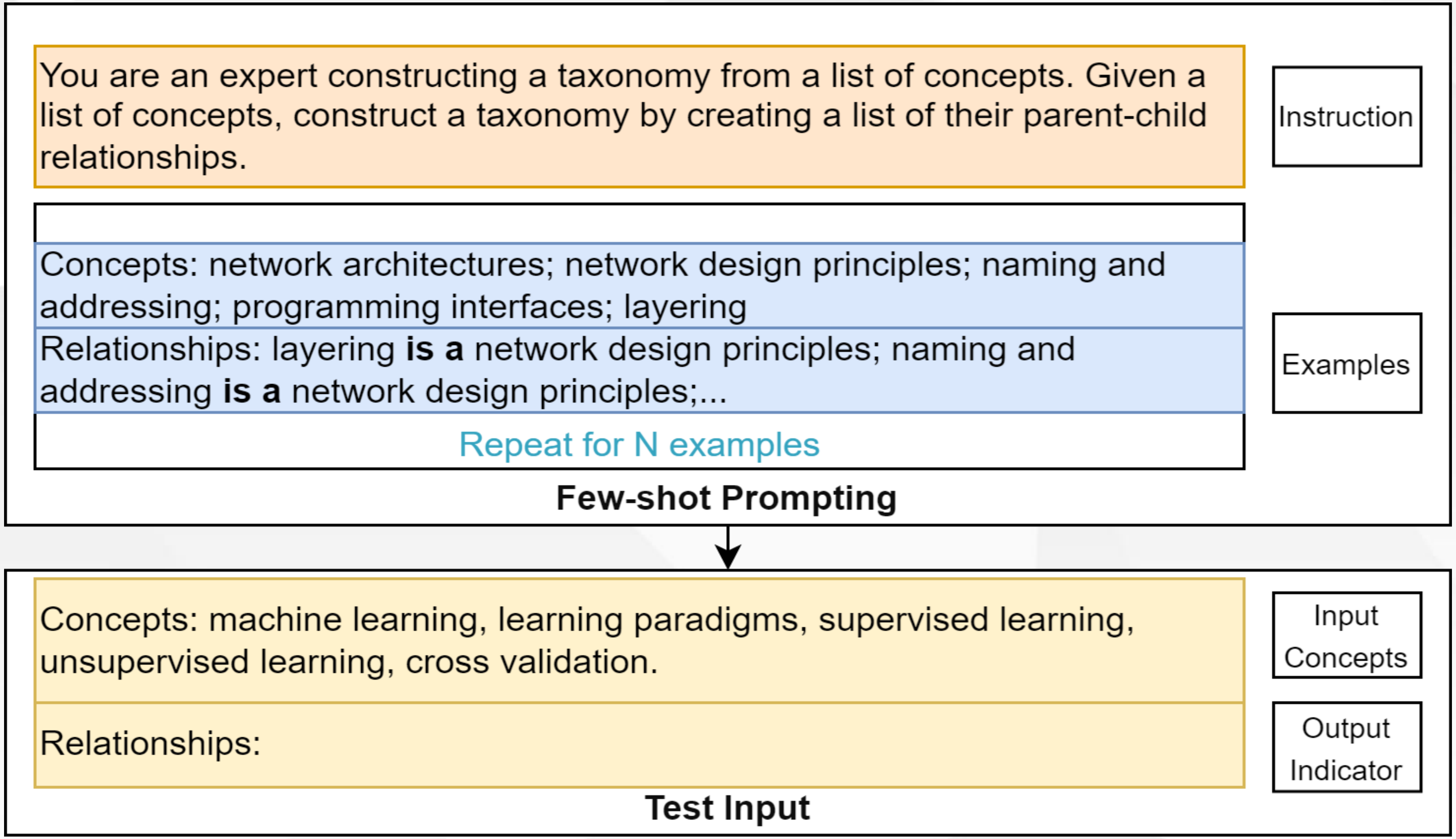
Test Input

Prompt

LLM

Relations: all have weight of 1

Relation Prediction: Prompt



You are an expert constructing a taxonomy from a list of concepts. Given a list of concepts, construct a taxonomy by creating a list of their parent-child relationships.

Instruction

Concepts: network architectures; network design principles; naming and addressing; programming interfaces; layering

Relationships: layering **is a** network design principles; naming and addressing **is a** network design principles;...

Repeat for N examples

Examples

Few-shot Prompting

Concepts: machine learning, learning paradigms, supervised learning, unsupervised learning, cross validation.

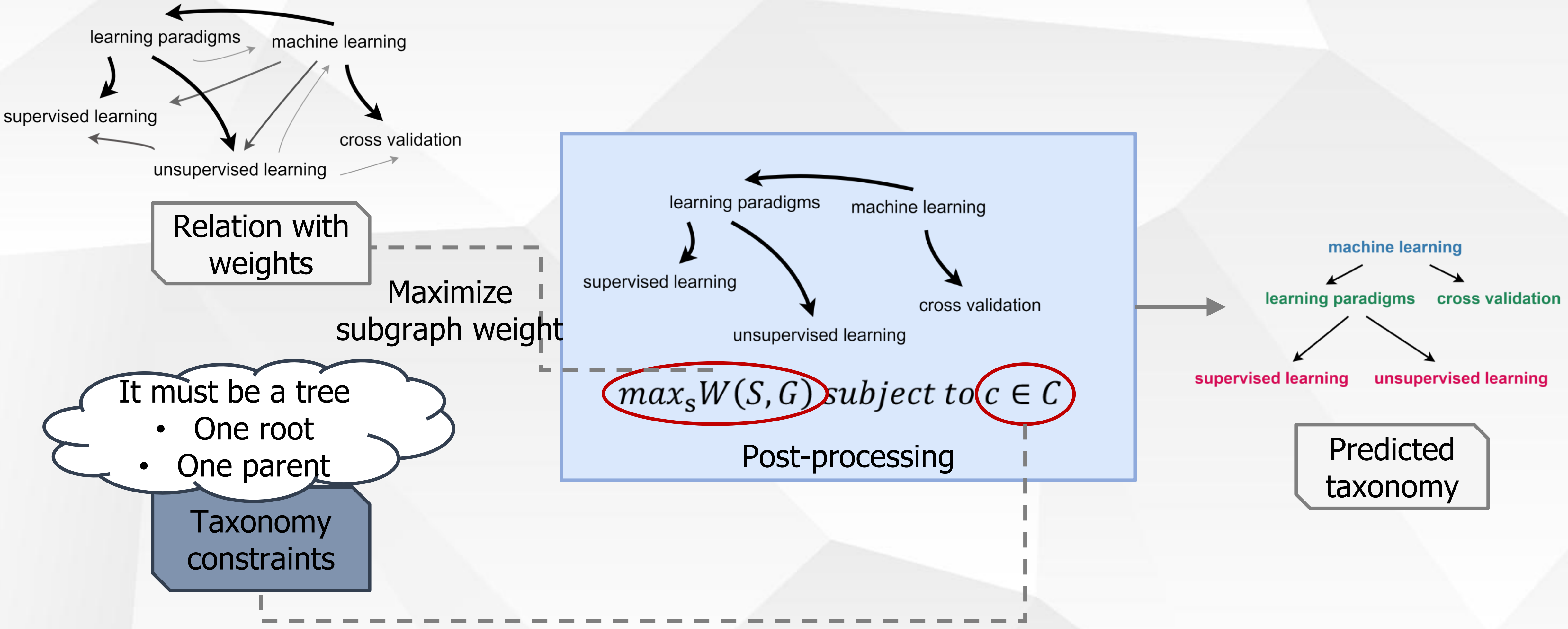
Input Concepts

Relationships:

Output Indicator

Test Input

Post-processing



learning paradigms

machine learning

supervised learning

cross validation

unsupervised learning

Relation with weights

Maximize subgraph weight

It must be a tree

- One root
- One parent

Taxonomy constraints

learning paradigms

machine learning

supervised learning

cross validation

unsupervised learning

$max_s W(S,G)$ subject to $c \in C$

Post-processing

machine learning

learning paradigms

cross validation

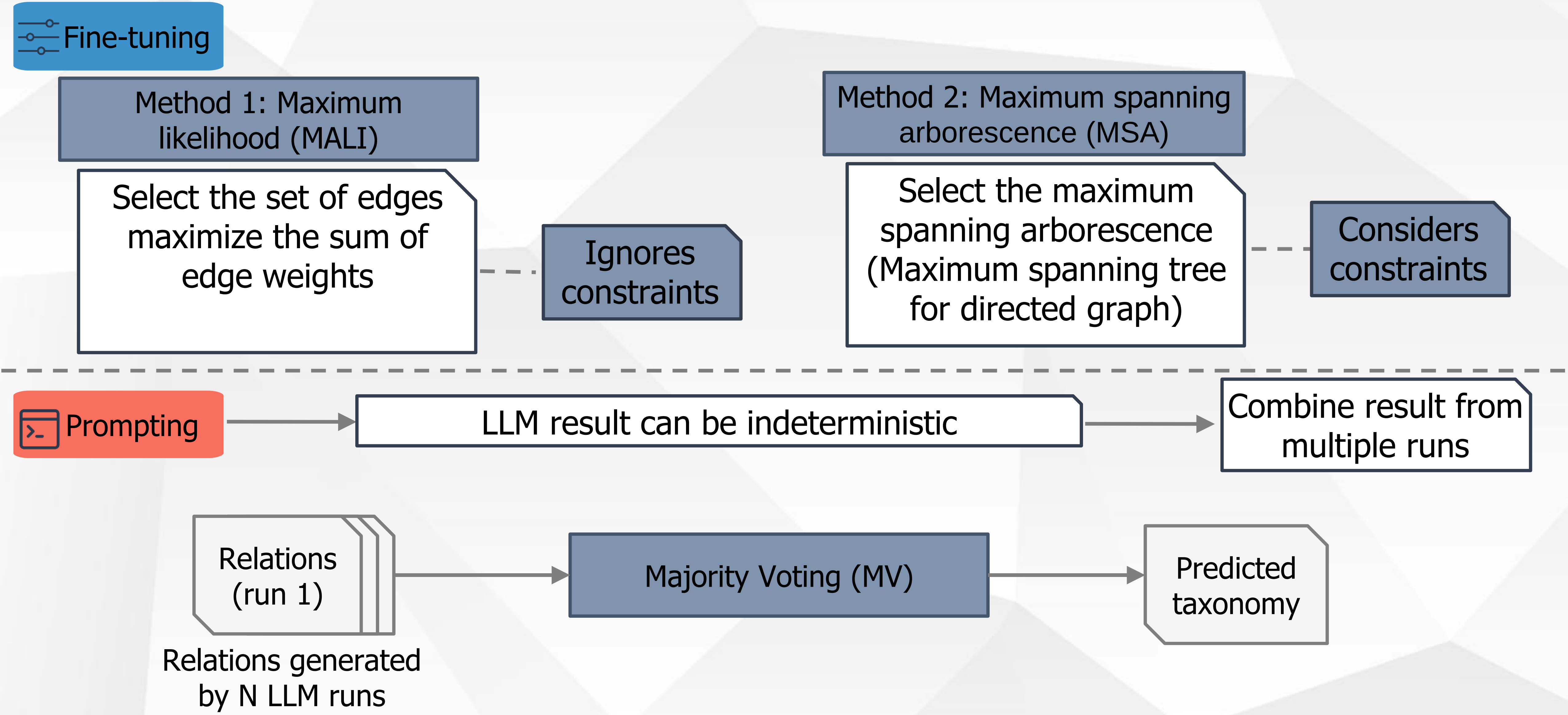
supervised learning

unsupervised learning

Predicted taxonomy

Respect the structural constraints

Post-processing



Fine-tuning

Method 1: Maximum likelihood (MALI)

Select the set of edges maximize the sum of edge weights

Ignores constraints

Method 2: Maximum spanning arborescence (MSA)

Select the maximum spanning arborescence (Maximum spanning tree for directed graph)

Considers constraints

Prompting

LLM result can be indeterministic

Combine result from multiple runs

Relations (run 1)

Relations generated by N LLM runs

Majority Voting (MV)

Predicted taxonomy

Research Questions:

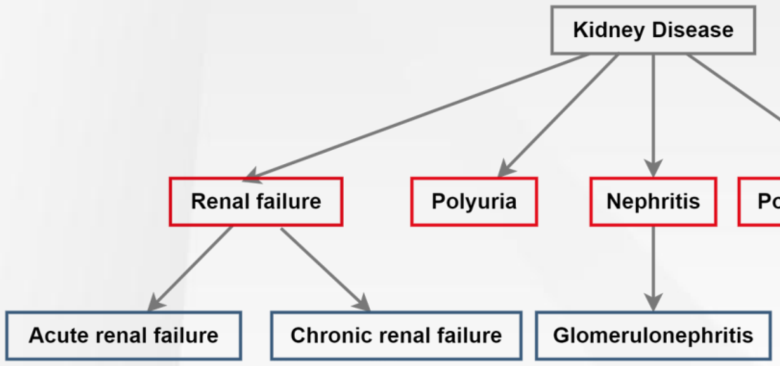
RQ1: How do the two LLM-based approaches differ when compared to the **ground truth**?

RQ2: What are the differences between the two LLM-based approaches in generating **consistent taxonomies**?

Dataset

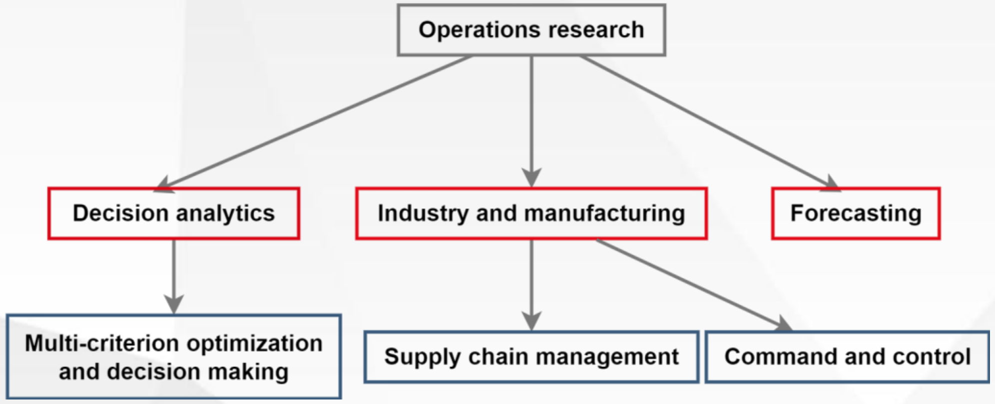
WordNet: A hypernym taxonomy (general English language concepts)

- **14,477** unique terms with **14,877 pairs**
- **761 taxonomies**
- **11 to 50** terms for each taxonomy



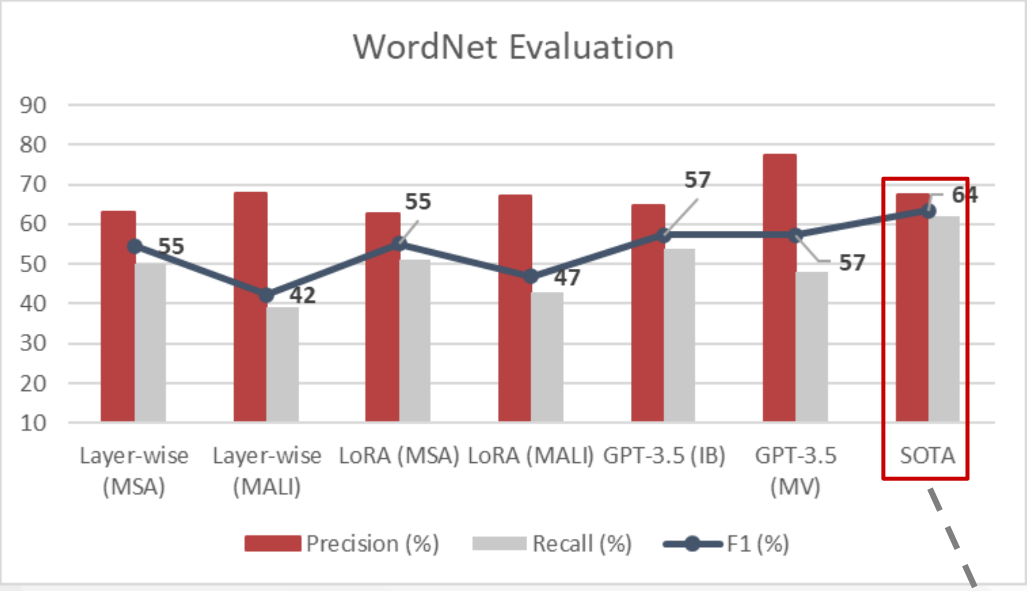
ACM CCS: newly created taxonomies in computer science derived from ACM Computing Classification System (CCS)

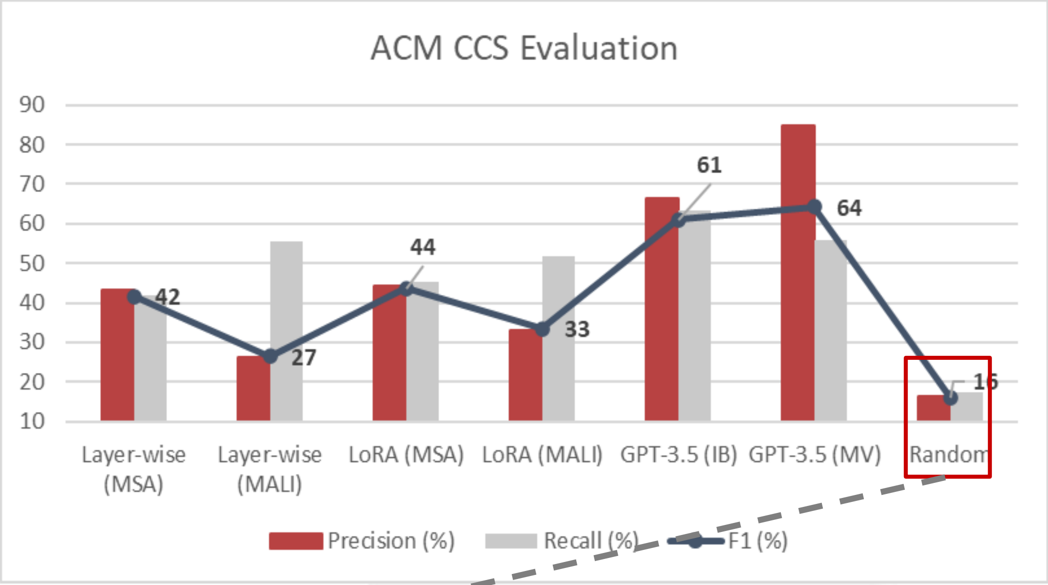
- **1846** unique terms with **1858 pairs**
- **75 taxonomies**
- **3 to 88** terms for each taxonomy



RQ1: Quality

RQ1: How do the two LLM-based approaches differ when compared to the **ground truth**?

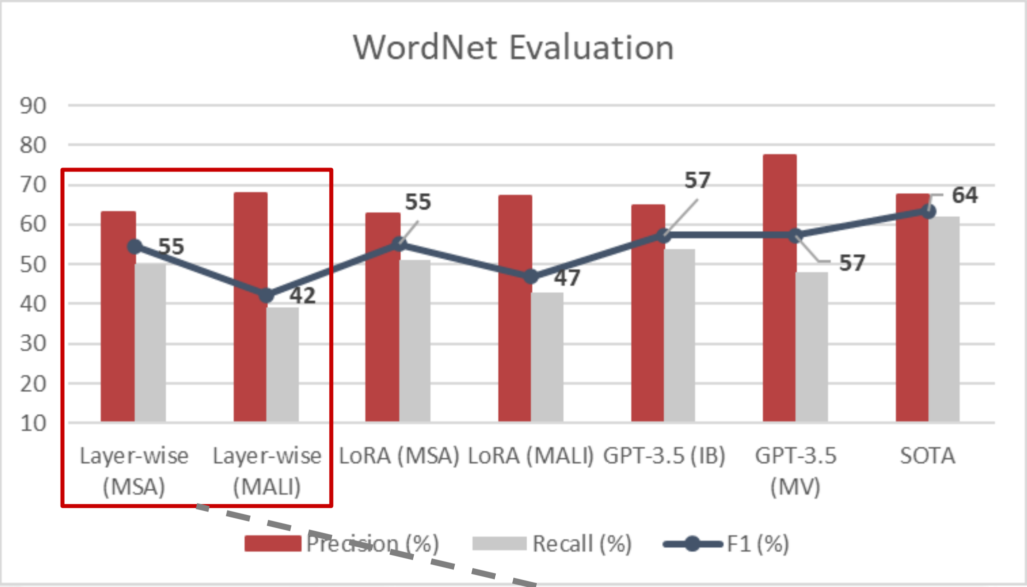


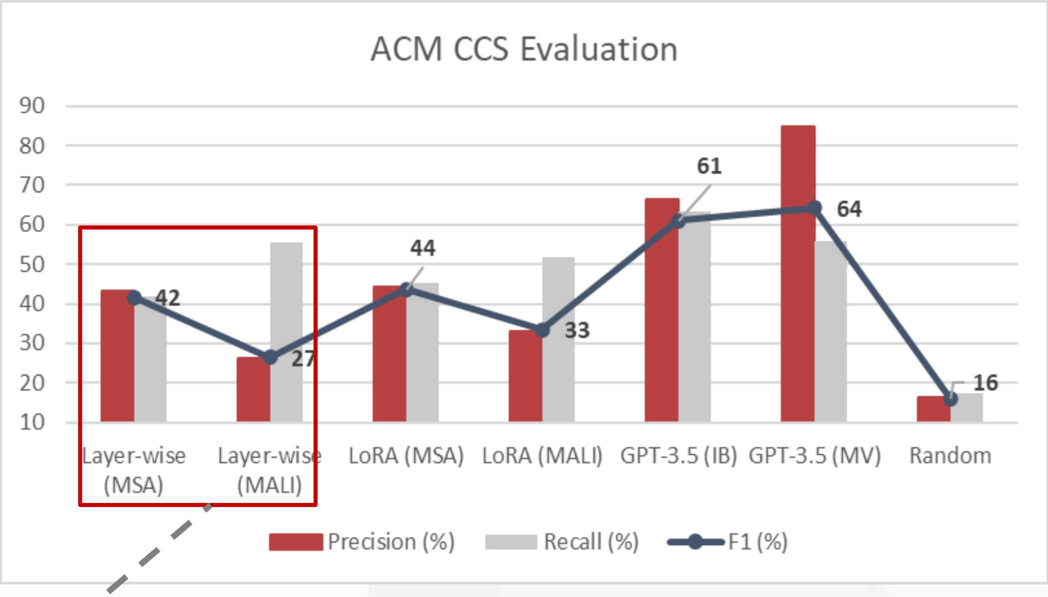


No methods beat SOTA, but all better than Random baseline

RQ1: Quality

RQ1: How do the two LLM-based approaches differ when compared to the **ground truth**?

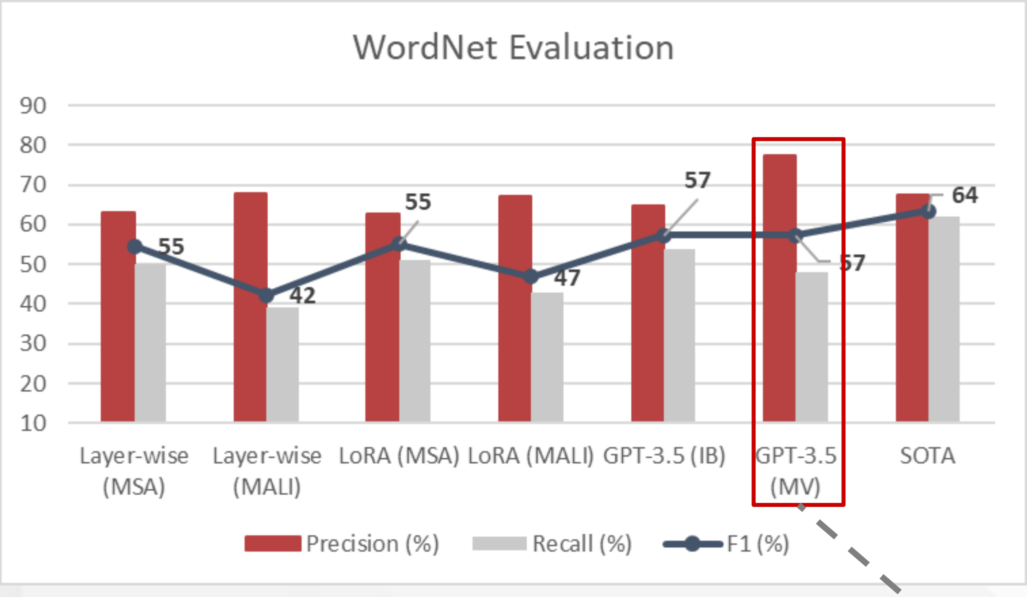


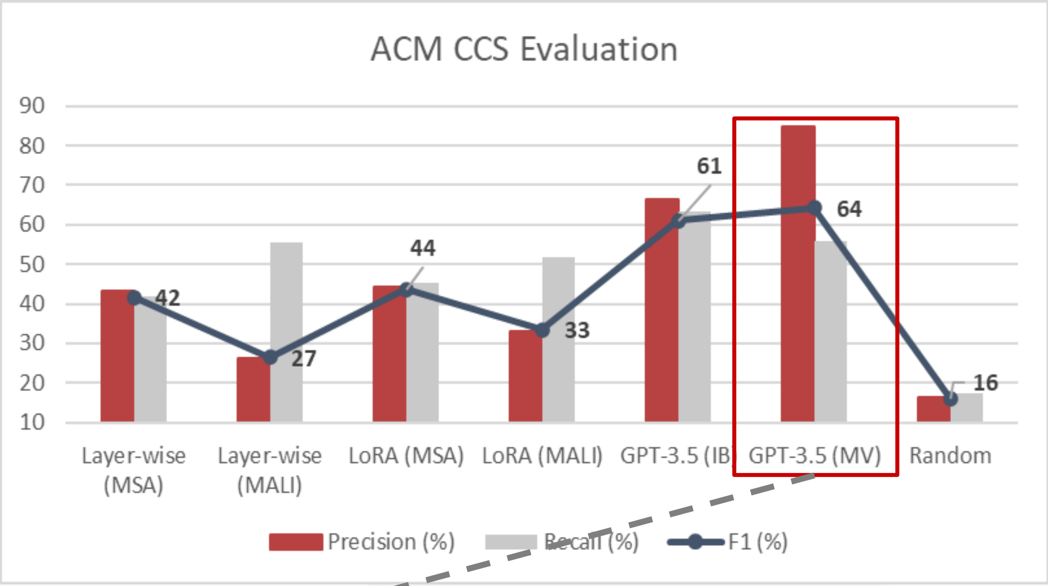


Improve consistency (MSA v.s. MALI) also improves the f1 score of the taxonomy

RQ1: Quality

RQ1: How do the two LLM-based approaches differ when compared to the **ground truth**?





Prompting is better than finetuning in both cases

RQ1: Quality

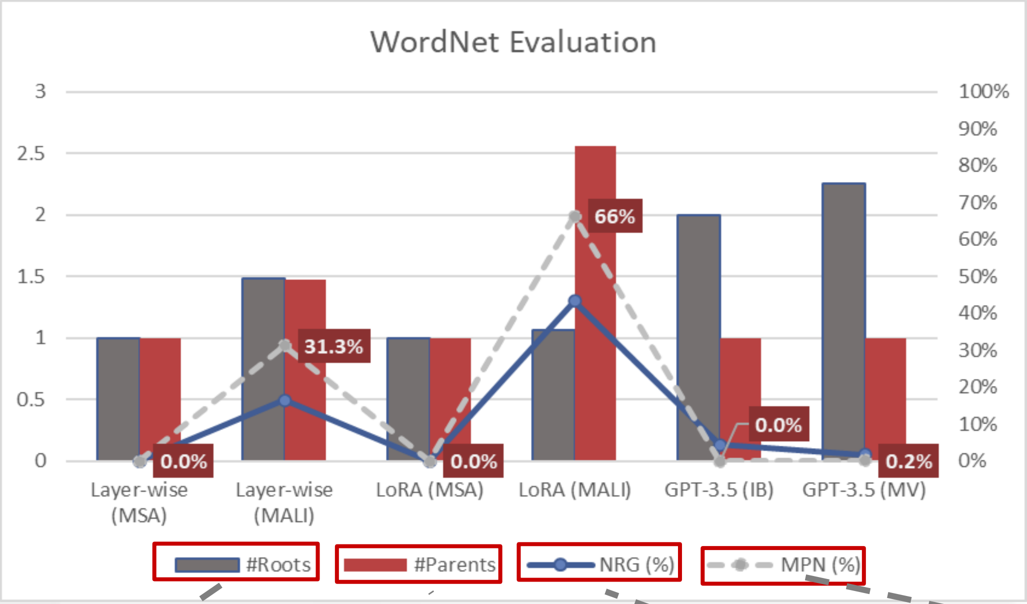
RQ1: How do the two LLM-based approaches differ when compared to the **ground truth**?

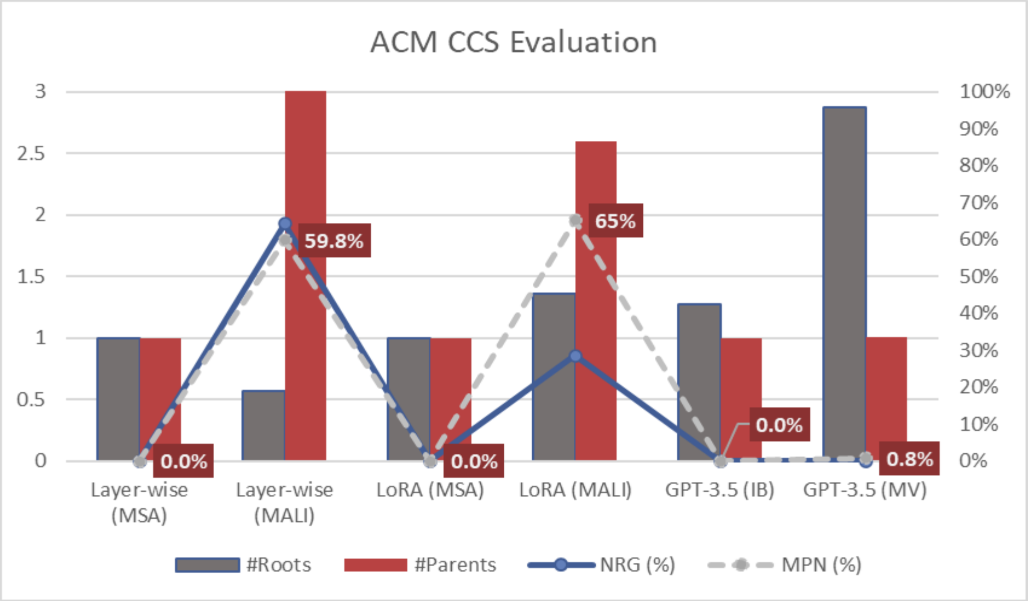
Answer:

- The **prompting method outperforms** the fine-tuning method in both datasets when comparing the **F1 and precision**.
- The performance **gap increases** when the **training dataset is smaller** (ACM CCS).

RQ2: Consistency

RQ2: What are the differences between the two LLM-based approaches in generating **consistent** taxonomies?





Number of roots of taxonomy (Consistent: 1)

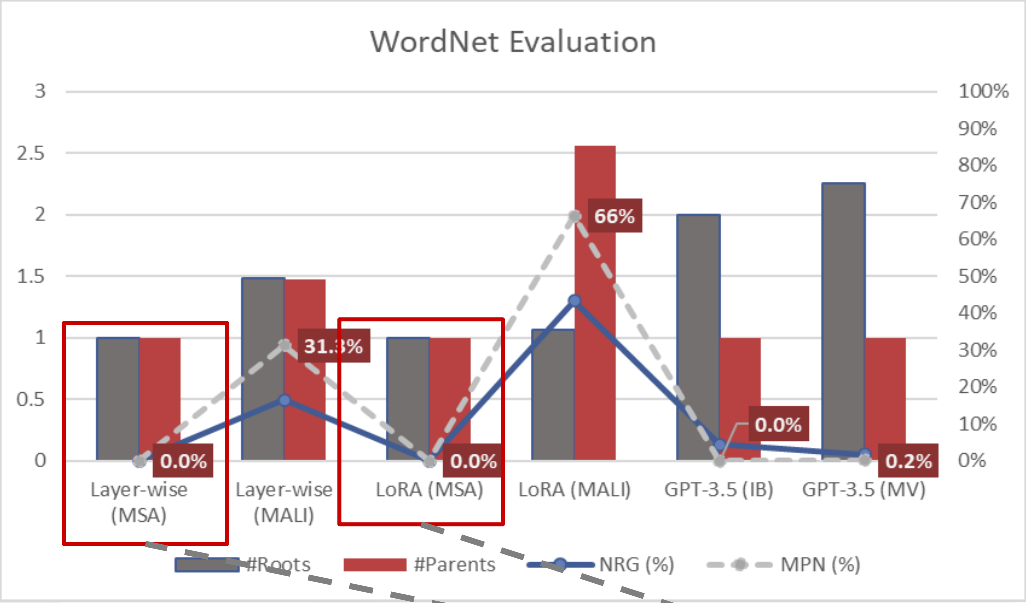
Number of parent for non-root nodes (Consistent: 1)

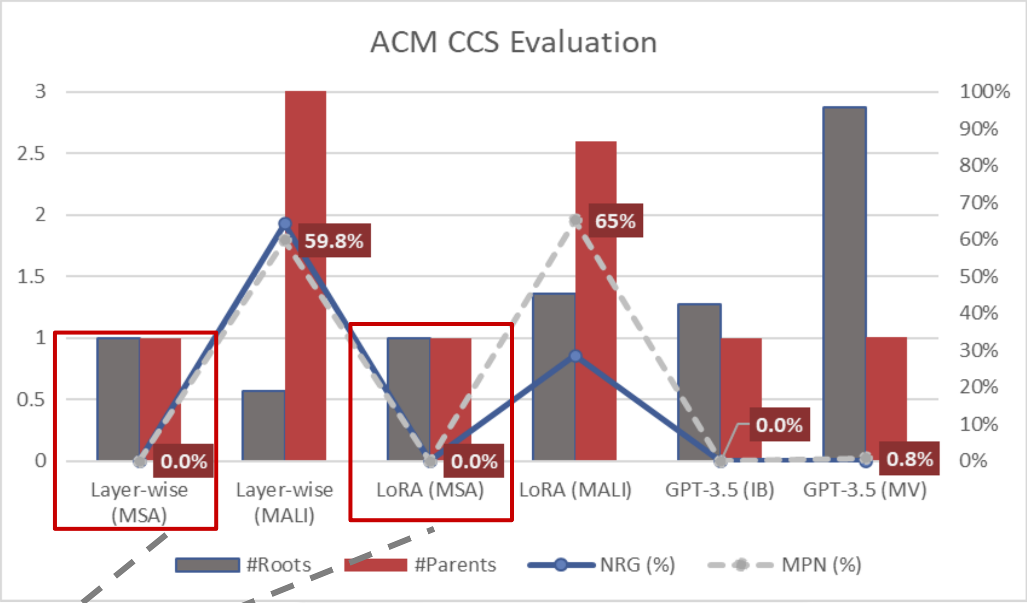
% of Taxonomies with no root (Consistent: 0)

% of non-root nodes with many parents (Consistent: 0)

RQ2: Consistency

RQ2: What are the differences between the two LLM-based approaches in generating **consistent** taxonomies?

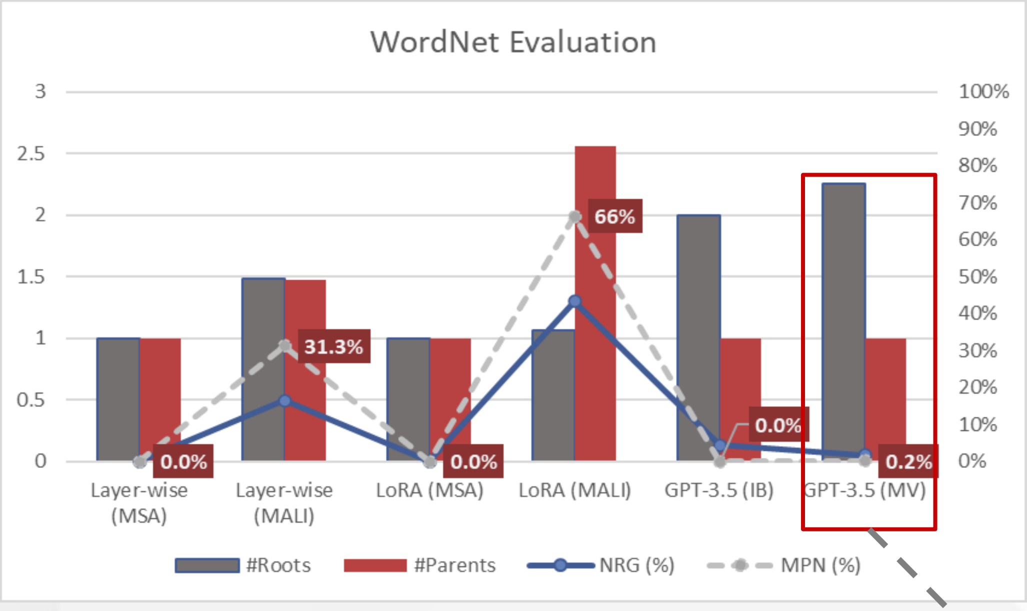

WordNet Evaluation

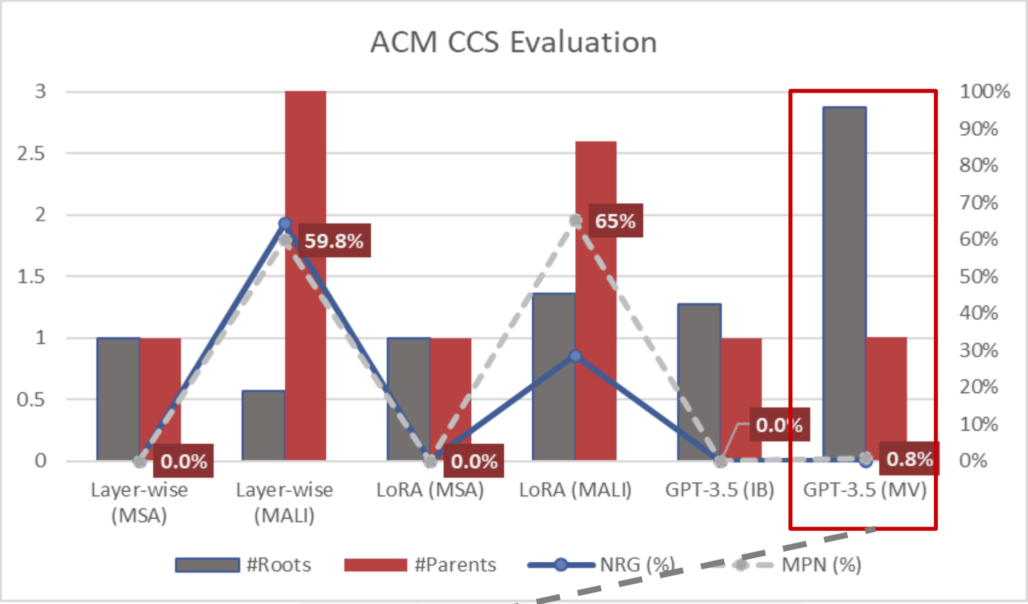

ACM CCS Evaluation

Maximum spanning arborescence achieves full consistency

RQ2: Consistency

RQ2: What are the differences between the two LLM-based approaches in generating **consistent** taxonomies?





Prompting with majority voting still contains some violations

RQ2: Consistency

RQ2: What are the differences between the two LLM-based approaches in generating **consistent** taxonomies?

Answer:

- Fine-tuning methods produce **fully consistent** taxonomies with the **MSA** post-processor.
- Taxonomies generated by the prompting approaches **still violate some constraints**

Discussion and Open Questions

Approach Selection: **Prompting** is a powerful tool and outperform finetuning

Taxonomy Consistency: LLM alone does not **guarantee consistency,** constraints need to be considered explicitly

Training data is not large enough

Tests appear in LLM's train data

Concept names are important

Automated Taxonomy Construction

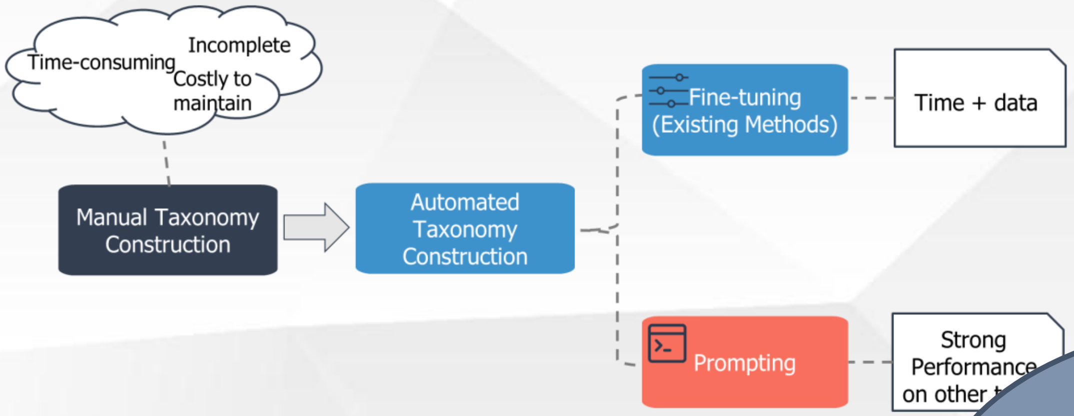
Combine post-processing and prompting

Extend to graph with multiple relation types

Extend to general constraints

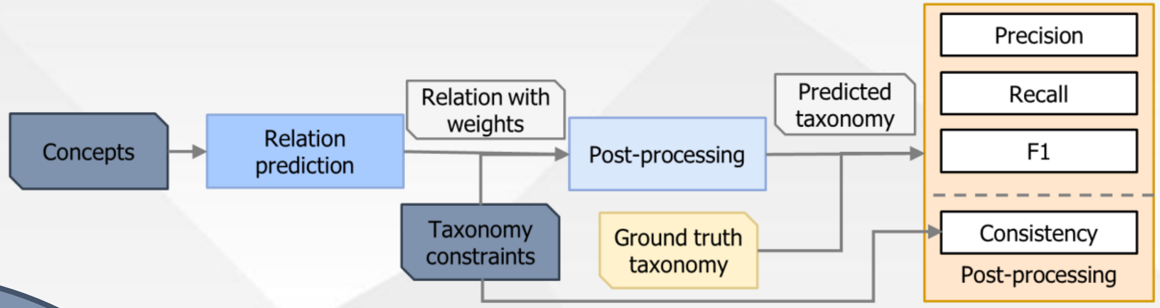
Conclusion

Motivation: Explore LLM for Taxonomy



Time-consuming
Incomplete
Costly to maintain
Manual Taxonomy Construction
Automated Taxonomy Construction
Fine-tuning (Existing Methods)
Time + data
Prompting
Strong Performance on other t

Approach Overview



Prompting or fine-tuning?

Evaluation

Research Questions:

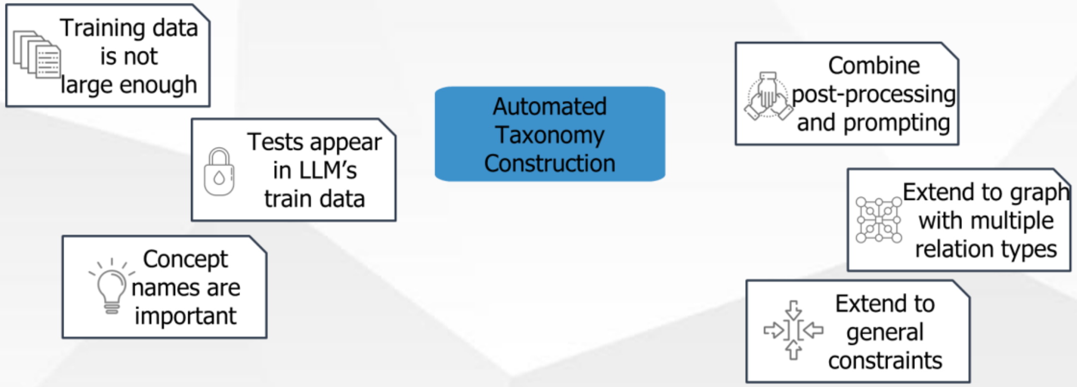
RQ1: How do the two LLM-based approaches differ when compared to the **ground truth**?

RQ2: What are the differences between the two LLM-based approaches in generating **consistent taxonomies**?

on and Open Questions

pproach Selection: **Prompting** is a powerful tool and outperform finetuning

Taxonomy Consistency: LLM alone does not **guarantee consistency**, but need to combine with *classic constraint optimization*.



Training data is not large enough
Tests appear in LLM's train data
Concept names are important
Automated Taxonomy Construction
Combine post-processing and prompting
Extend to graph with multiple relation types
Extend to general constraints