Cofinanțat de
Uniunea Europeană

HUB Român de Inteligență Artificială
cod SMIS: 351416

Knowledge Graph Integration in Natural Language Processing for the Smart Cooking Domain

Vlad VĂDEAN
Vlad NEGRU
Camelia LEMNARU
Rodica POTOLEA

Workshop HRIA - Intelligent and Autonomous Systems 2

06/11/25

UNIVERSITATEA TEHNICĂ
DIN CLUJ-NAPOCA

Problem Context

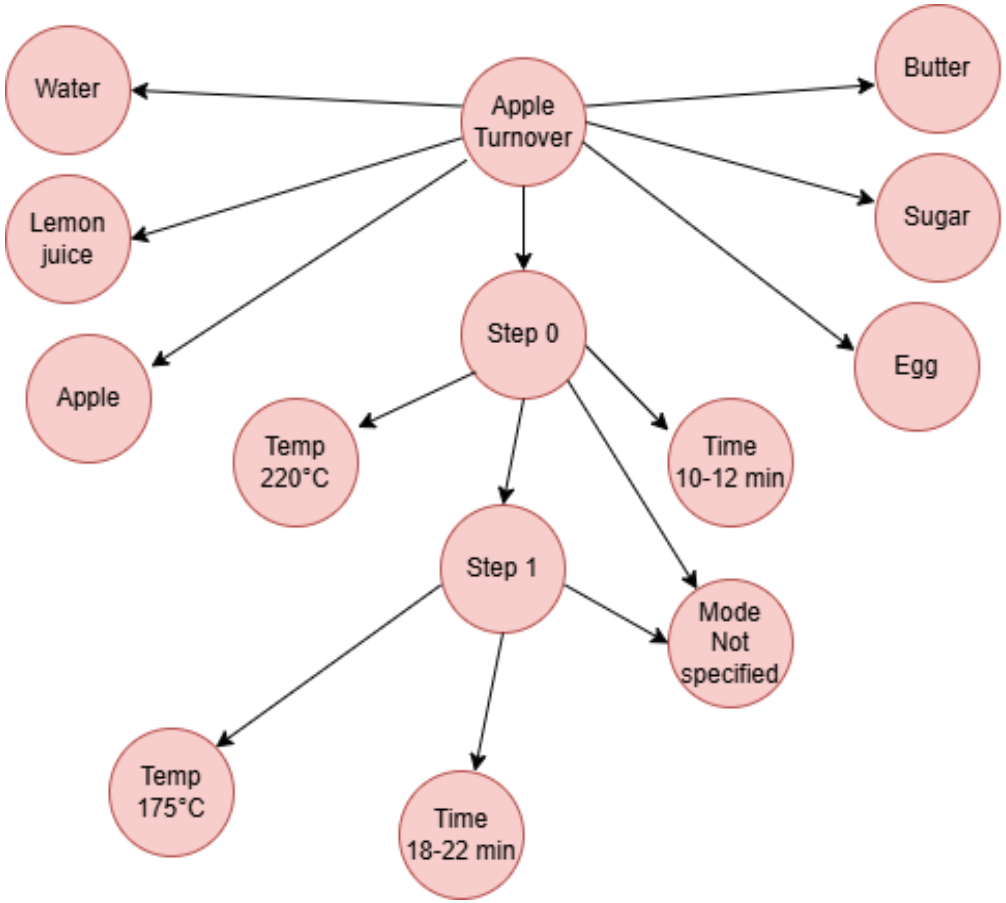
APPLE TURNOVERS

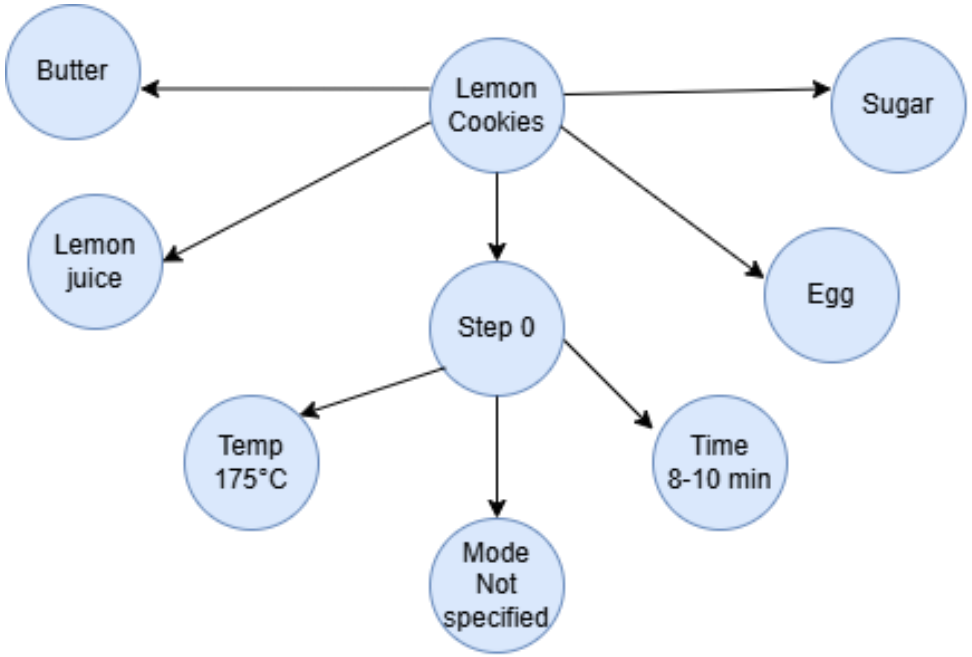
In a small bowl, combine flour and salt; cut in butter until crumbly. Gradually add water, tossing with a fork until a ball forms. Divide dough in half; wrap in plastic and refrigerate for 30 minutes. Preheat oven to 400°. In a medium saucepan, combine sugar, cornstarch, cinnamon, and nutmeg. Add apples and lemon juice; toss to coat. Bring to a boil over medium heat, stirring constantly. Reduce heat; simmer, uncovered, until apples are tender, stirring often, 8-10 minutes. Remove from heat and let cool slightly. In a small bowl, whisk together egg and water. On a lightly floured surface, roll one dough half into a 10-in. square; cut into four squares. Brush edges with egg mixture. Spoon about 2 tablespoons filling onto center of each square; fold diagonally to form triangles. Press edges with a fork to seal. Place on a parchment-lined baking sheet. Repeat with remaining dough and filling. Brush turnovers with remaining egg mixture. Cut two small slits in the top of each turnover. Bake until golden brown, 18-22 minutes. Remove to a wire rack.

LEMON COOKIES

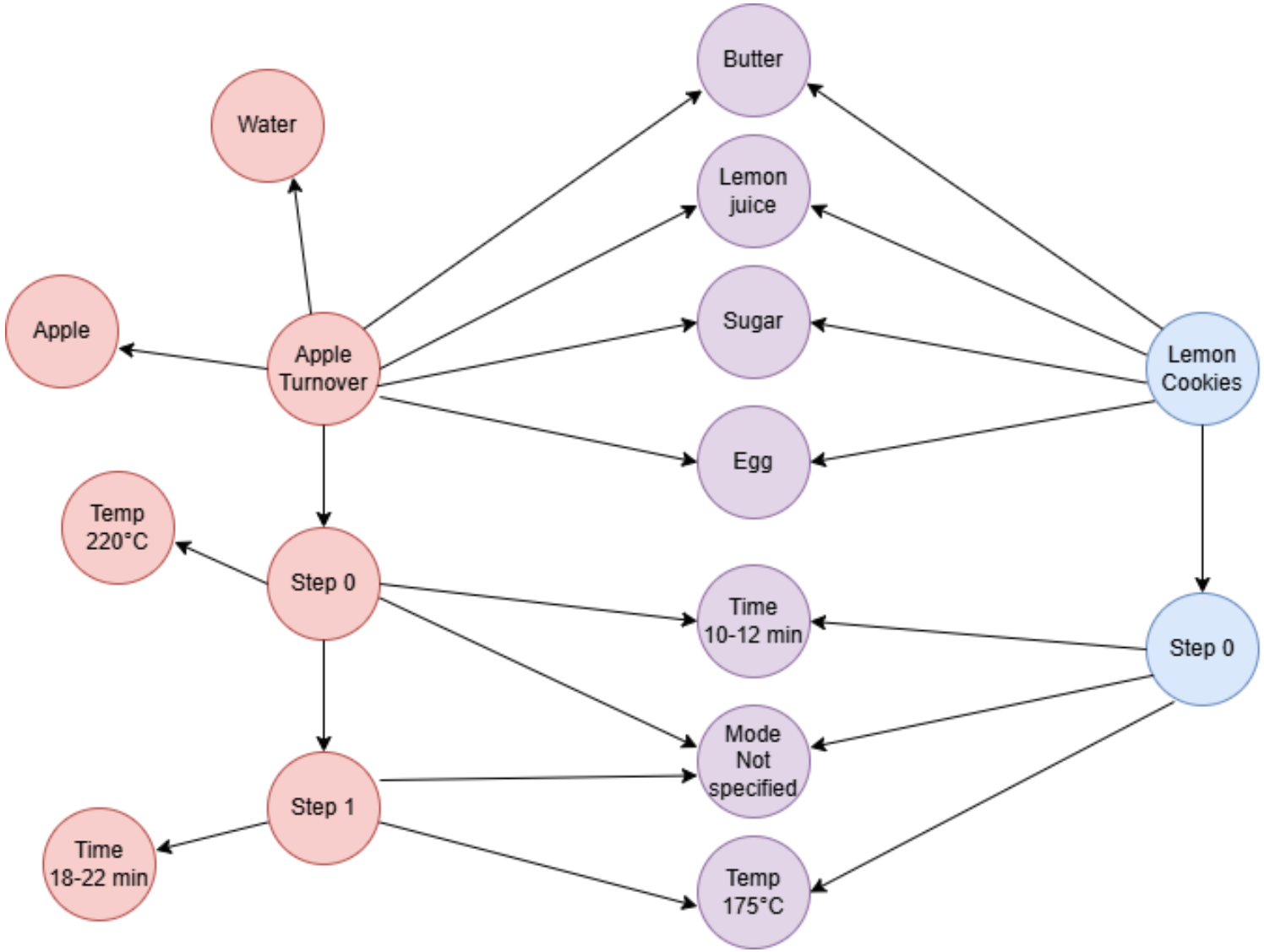
In a medium bowl, whisk together flour, baking powder, and salt. In a large bowl, beat butter and sugar until light and fluffy. Add eggs one at a time, beating well after each addition. Beat in lemon zest, lemon juice, and vanilla extract. Gradually add flour mixture to butter mixture, stirring until just combined. Cover bowl with plastic wrap; refrigerate for 30 minutes. Preheat oven to 350°. Line baking sheets with parchment paper. Scoop dough by tablespoonfuls and roll into balls. Place on prepared baking sheets, spacing 2 inches apart. Flatten slightly with the bottom of a glass. Bake until edges are lightly golden, 8-10 minutes. Remove to a wire rack and cool for 5 minutes. Meanwhile, in a small bowl, whisk together powdered sugar, lemon juice, and lemon zest until smooth. Drizzle glaze over warm cookies. Let stand until glaze sets, about 15 minutes. Serve at room temperature.

Problem Context





Problem Context



Problem Context

TASK:

Given a multilingual recipe dataset, find the most similar recipe to a user's text description.

Starting QUESTION:

How to compare multiple entities and measure their similarity?

- Behavior-based equivalence
    - used in an observable dynamic context
- Feature-based equivalence
    - used in a well-known predefined feature space

Problem Context

Feature-based equivalence

Main QUESTIONS:

1. How to select the relevant features to underline the differences between distinct entities?

2. How can we organize features to enable easy computation of a similarity metric & reconstruct the entity corpus?

Exploring Possibilities

1. How to select the relevant features to underline the differences between distinct entities?

Attributes should be informative about the entity, not merely correlated with it!

- structural attributes: components, processes, steps, ingredients
- summarization attributes: ranking, goal, grade, quality, difficulty
- identity attributes: role, categories, types, names, titles

Exploring Possibilities

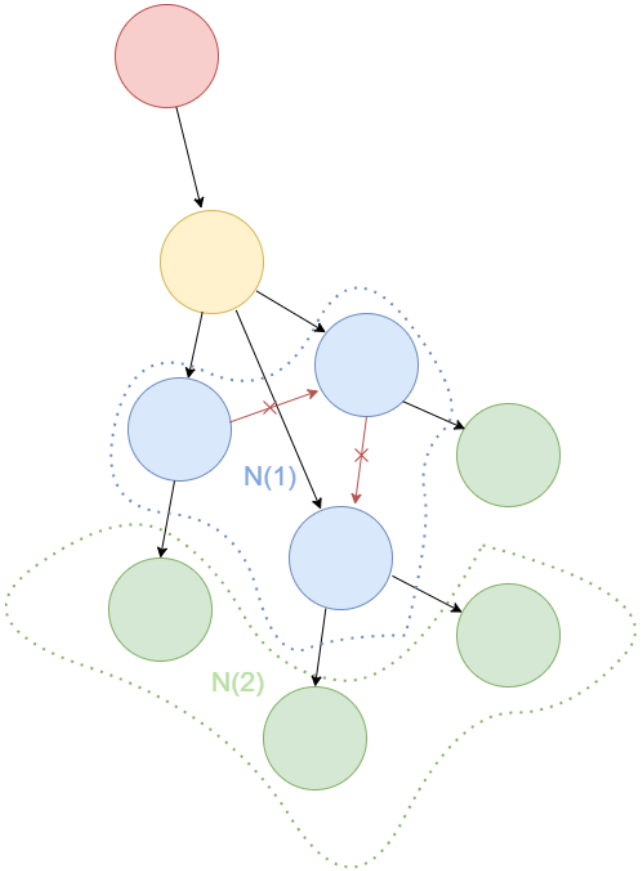
2. How can we organize features to enable easy computation of a similarity metric & reconstruct the entity corpus?

Data representation

- embedding
- encoding

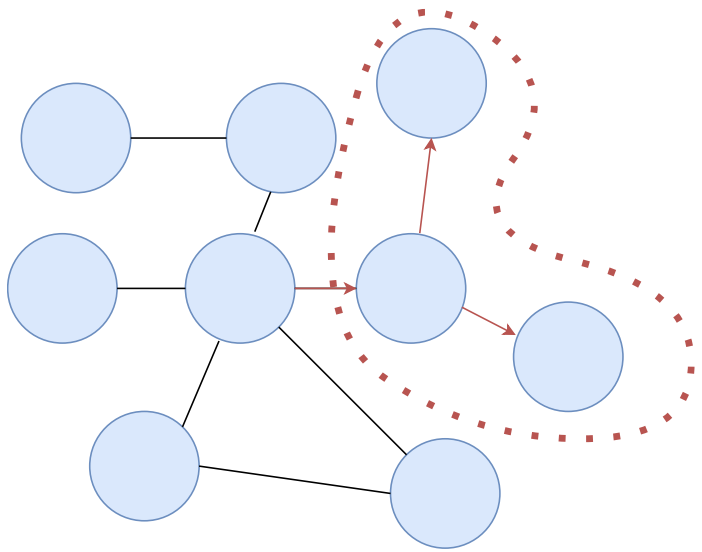
Data linkage

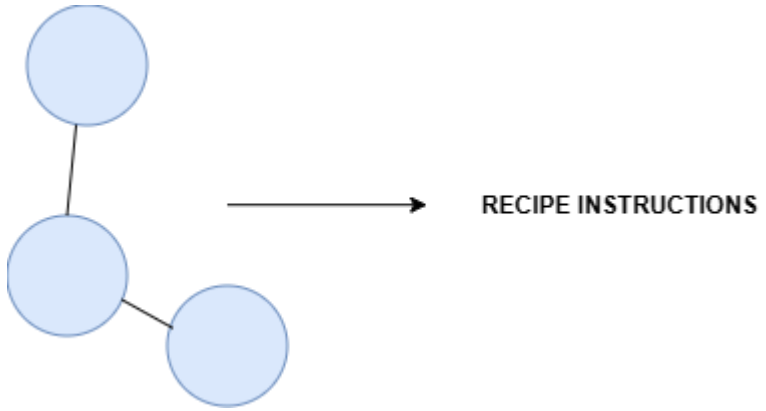
- graph-based ontology



Objectives

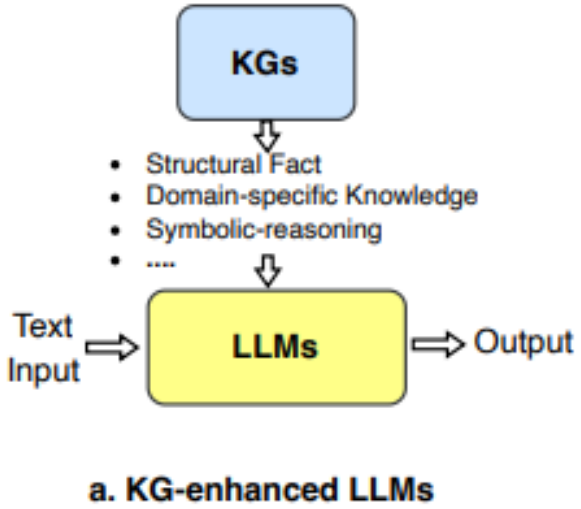
1. Build a **bottom-up knowledge graph**.
2. Assign **proper values** to the data.
3. Extract **relevant features** and build the input recipe subgraph.
4. Find the most **suitable candidate**, based on **similarity measurements**.
5. Reconstruct the **entity corpus**.



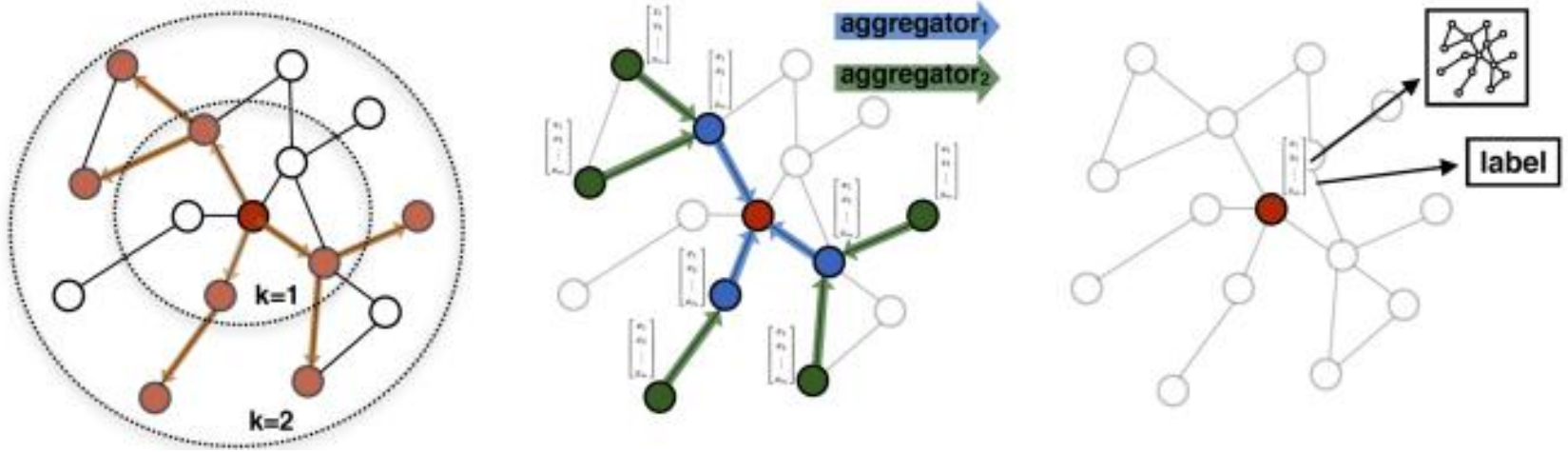


Related work

- KG-enhanced LLM Architecture [1]
- GraphSAGE [3]



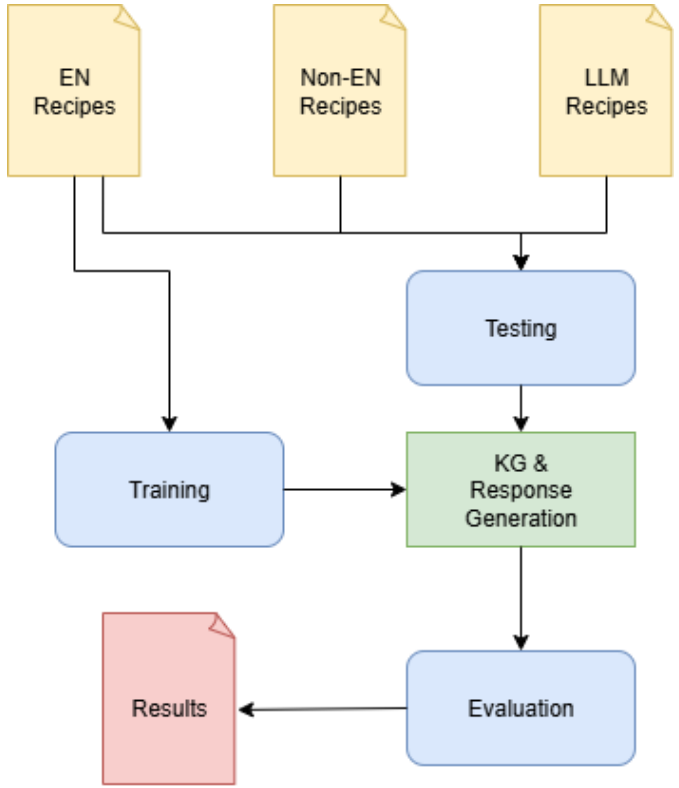
a. KG-enhanced LLMs



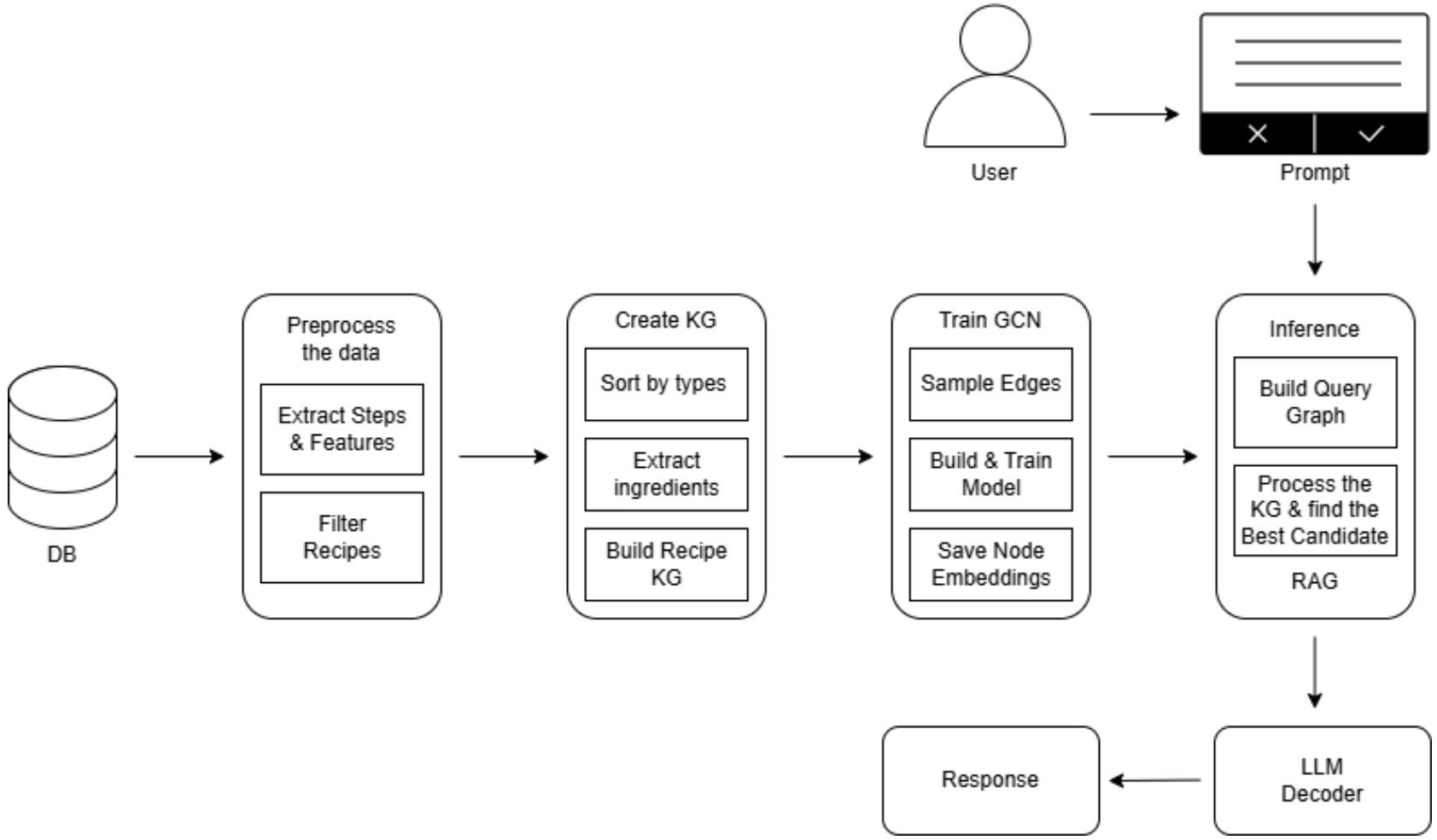
Dataset

- The dataset is multilingual and contains the title, instructions and properties of the cooking recipes

	Avg. No. Ingredients	Max - Min No. Ingredients	Avg. No. Steps	Max - Min No. Steps	No. Recipes
fr	0.42	7 - 0	1.08	5 - 0	2749
de	0.74	9 - 0	1.10	6 - 0	28250
it	1.45	8 - 0	1.12	5 - 0	2003
nl	0.93	9 - 0	1.13	6 - 0	6137
sv	0.22	6 - 0	1.12	4 - 0	749
en	7.88	34 - 0	1.18	5 - 0	12042

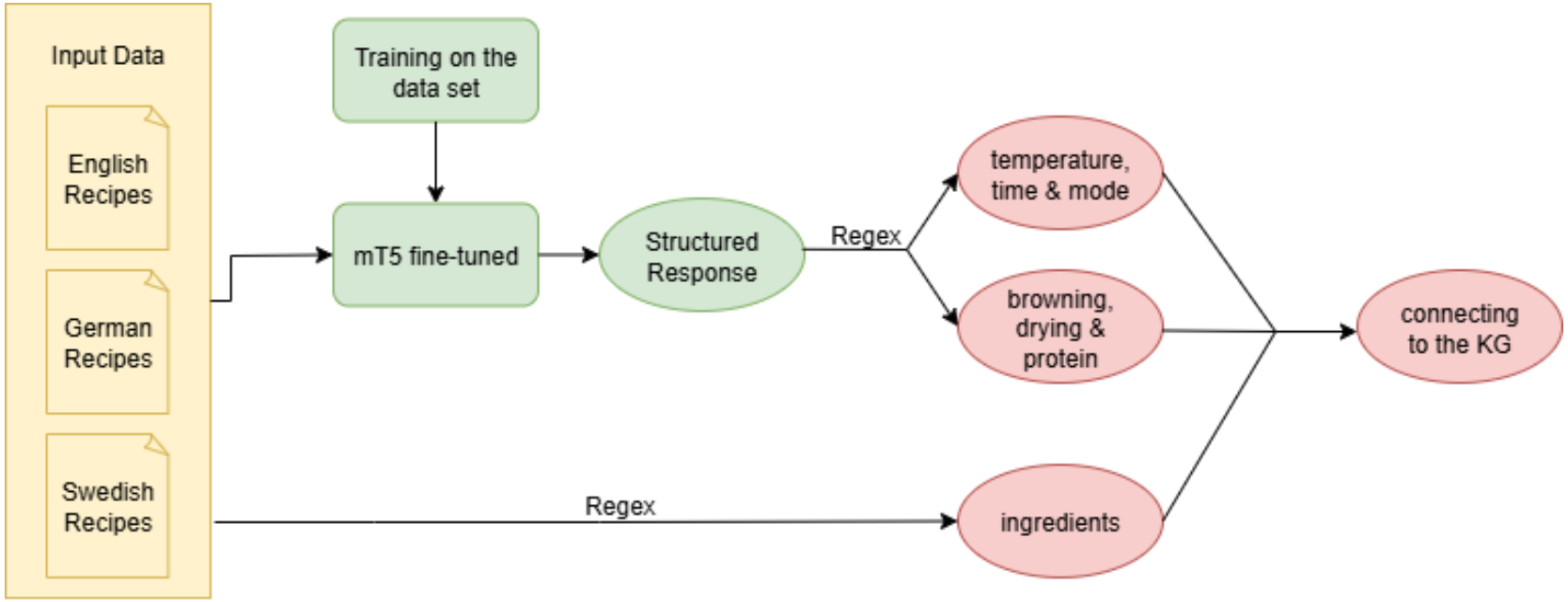


Proposed Solution – Conceptual Architecture



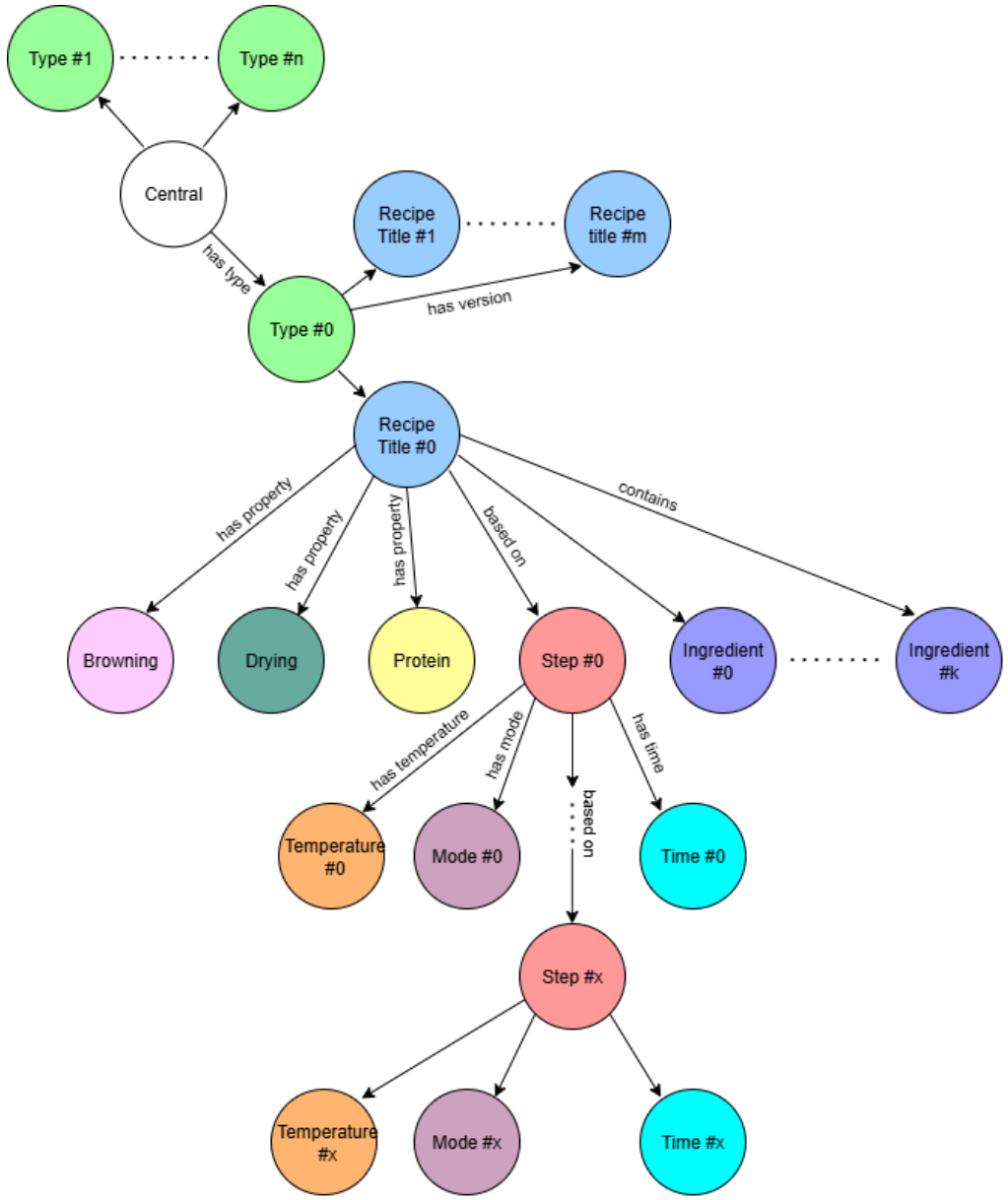
Proposed Solution – Building Recipe Subgraph

- Using a fine-tuned mT5 Base model to solve multilingual data extraction
- To extract ingredients, we apply regular expressions (regexes) to the user's prompt.



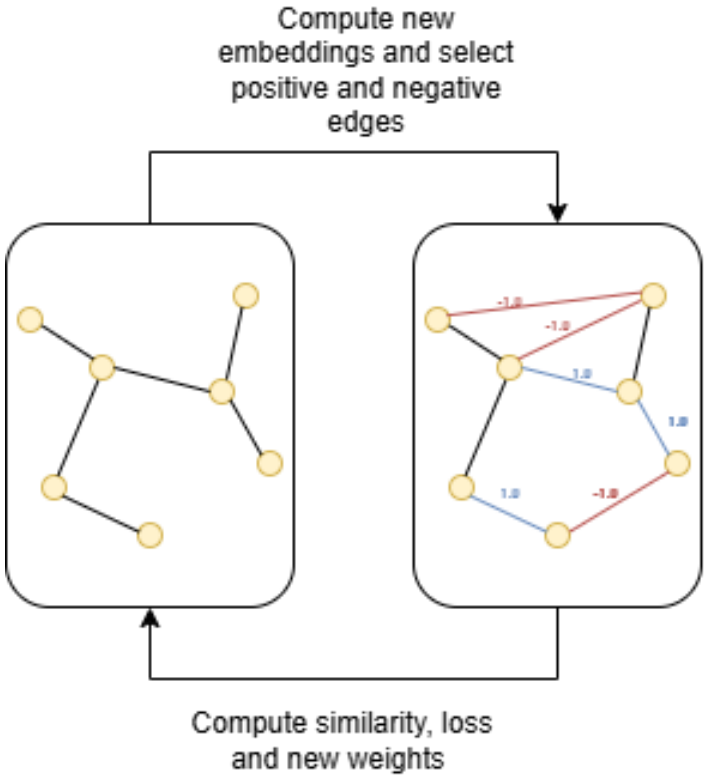
Proposed Solution - KG

- The implemented structure is a DAG to avoid contradictions and to enforce a semantic hierarchy.
- Network construction starts from the recipe subgraphs, which are linked via type nodes and then connected through the central node.
- Nodes - textual value and type.
- Edges - labeled to represent different semantic relations.
- 171,300 edges; 22,215 nodes; 57 types; 9,690 recipes.



Proposed Solution - GCN

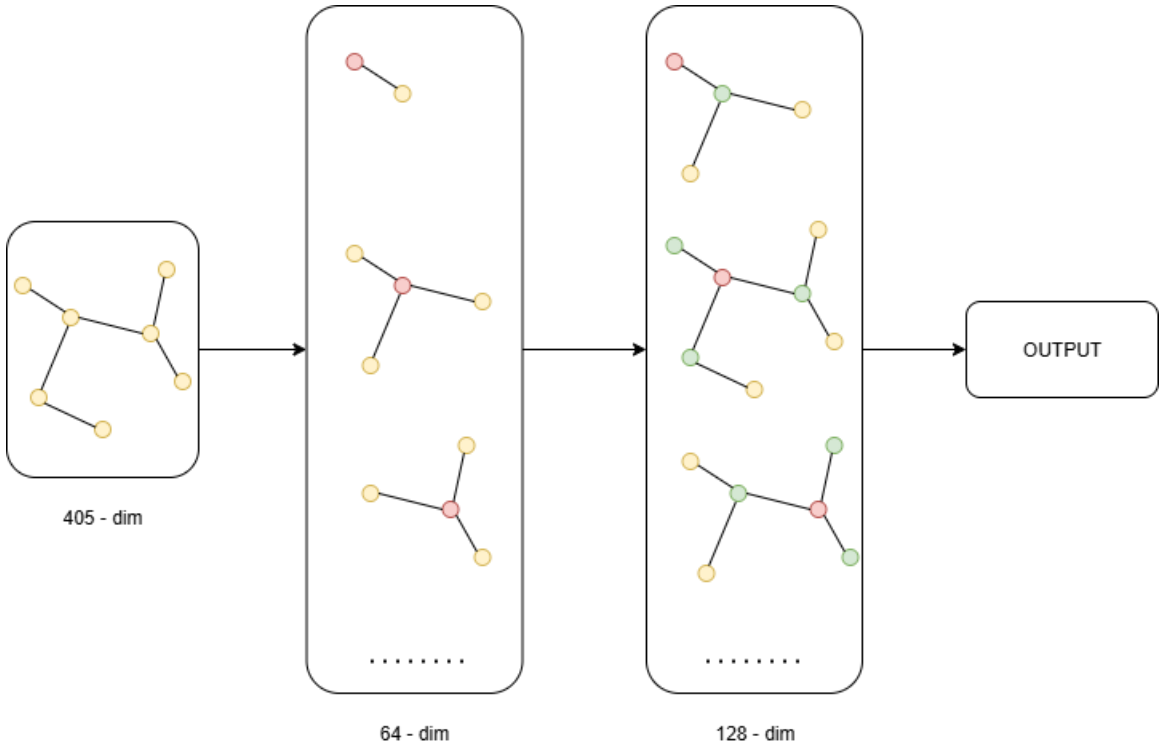
- GCN – Graph Convolutional Network
- The GCN is used for creating meaningful embeddings for each node.
- The loss function is trying to force the GCN to assign more closer embeddings for the nodes that have a direct edge then the ones that do not.



$$\text{loss} = \text{ReLU}\Big(\text{margin} - \big(\text{sim_pos.mean}() - \text{sim_neg.mean}()\big)\Big)$$

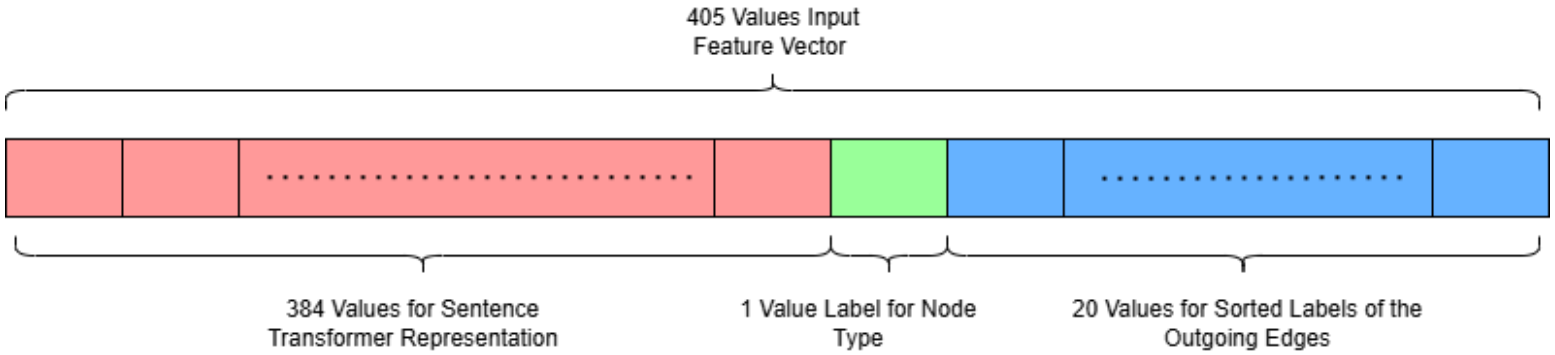
Proposed Solution - GCN

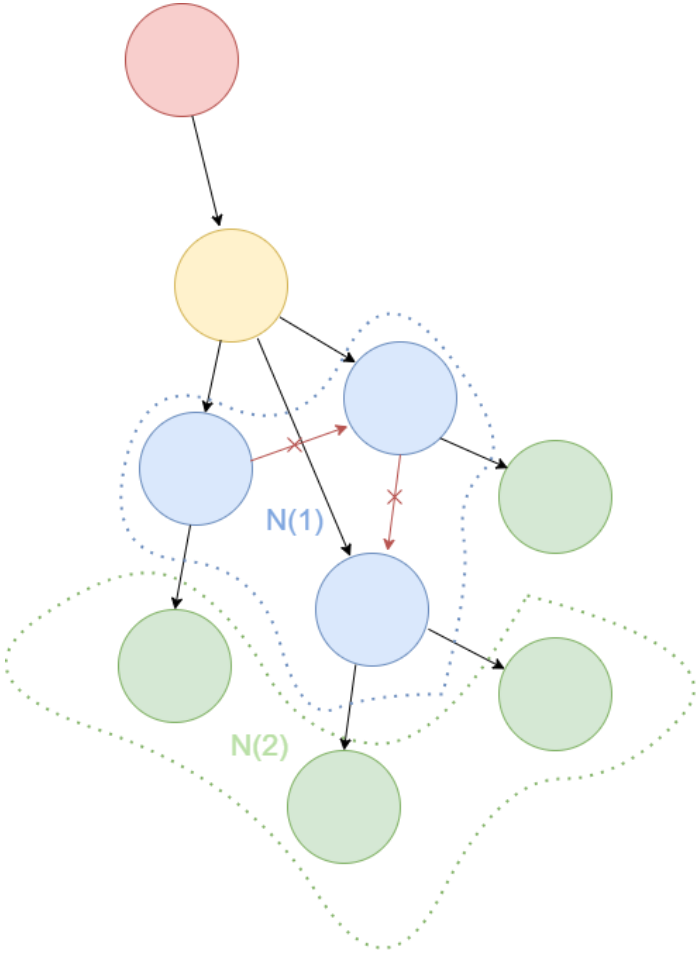
- To support inference for recipes containing nodes not seen during training, we use an inductive model, similar to GraphSAGE [3].
- Input layer – 405, hidden layer – 64, output layer – 128
- The 64 layer uses a ReLU, to add non-linearity
- The mean aggregator is used to allow the elements to participate equally to the final result.



Proposed Solution - GCN

- 405 embedding is the input vector made of:
  - 384 embedding of the text value of the node
  - 1 value for the encoded node type
  - 20 sorted values for the encoded labels of the 1-hop outgoing edges of the node





Proposed Solution – Building Recipe Subgraph

INDUCTIVE:

- learns **how to generate** embeddings from features
- generalizes to **newer, unseen nodes**
- **does NOT require** retraining if new nodes are added
- **Graph SAGE**
- $H_v^k = \sigma[W \frac{1}{1+|N(v)|}(H_v^{k-1} + \sum_{u \in N(v)} H_u^{k-1})]$

TRANSDUCTIVE:

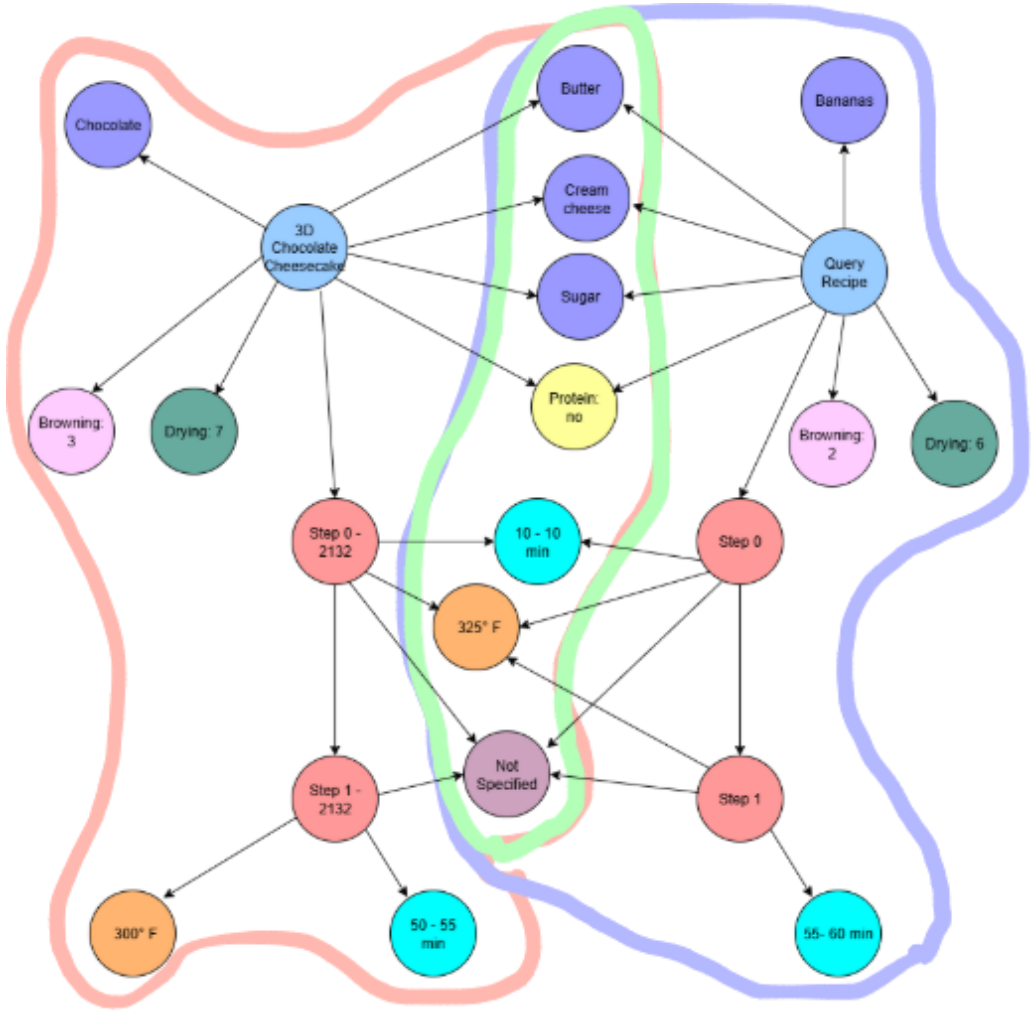
- learns **fixed embeddings** for specific nodes
- requires **complete graphs** during training
- **does require** retraining if new nodes are added
- **Classic GCN**

$$H^k = \sigma(D^{\frac{-1}{2}} A D^{\frac{-1}{2}} H^{k-1} W)$$

Proposed Solution – Similarity

- Similarity is computed by grouping nodes of the same type.

- To avoid selecting a candidate where the query graph is neither equivalent to nor contained in it, we will use the ratio of shared ingredients and shared nodes.



Results – Similarity Formulas

INITIAL_SCORE = PROPERTY_SIMILARITY + STEP_SIMILARITY + INGREDIENT_SIMILARITY

COMMON_NODES_SCORE = PROPERTY_SIMILARITY + STEP_SIMILARITY + INGREDIENT_SIMILARITY * (NO_COMMON_INGREDIENTS / NO_TOTAL_INGREDIENTS)

COMMON_RELEVANT_SCORE = COMMON NODE SCORE* (NO_COMMON_NODES / NO_TOTAL_NODES)

	Correct	Err < 0.1	Err >= 0.1	N/A	Multiple
INITIAL	77.56	20.34	1.14	**0.96**	-
COMMON NODES	97.80	0.55	**0.47**	**0.96**	0.20
COMMON RELEVANT	**97.91**	**0.42**	0.55	**0.96**	**0.17**

Results – Inductive GCN vs Transductive GCN

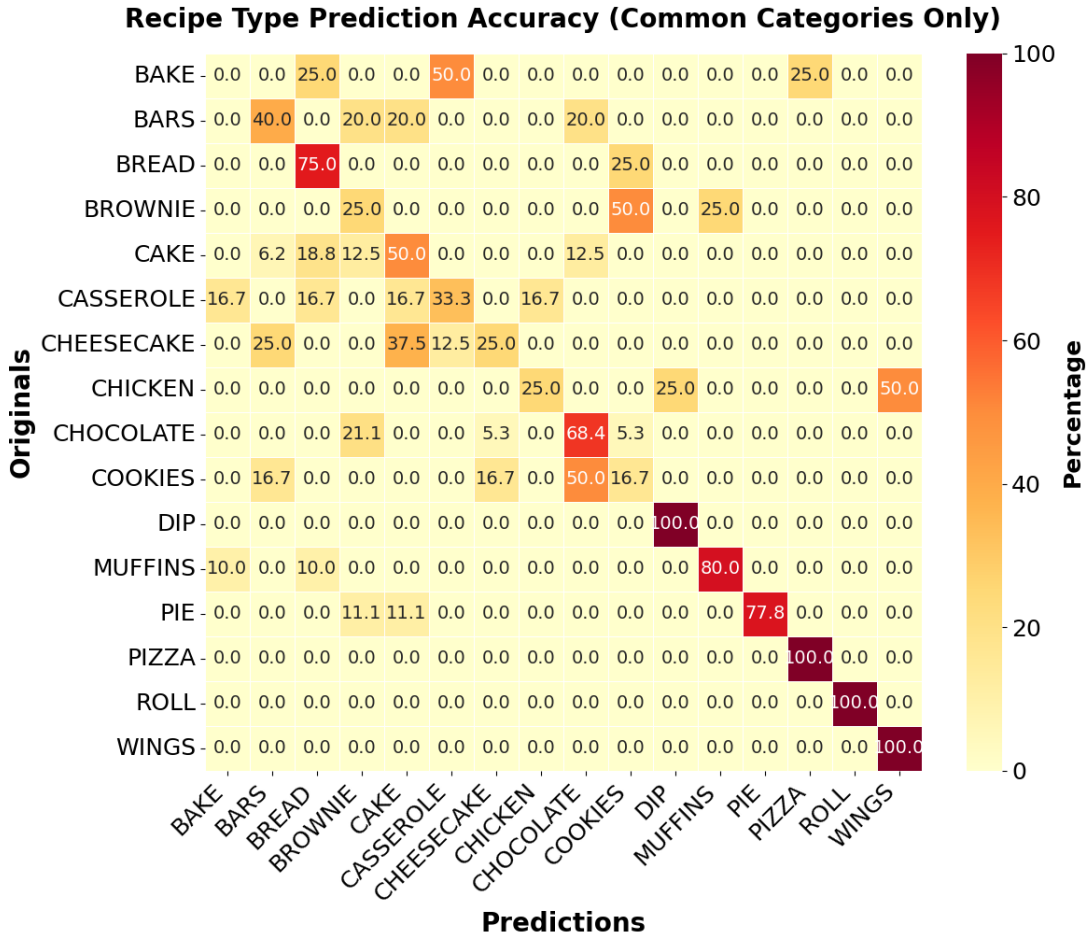
- Average similarity on successfully found recipes from the database

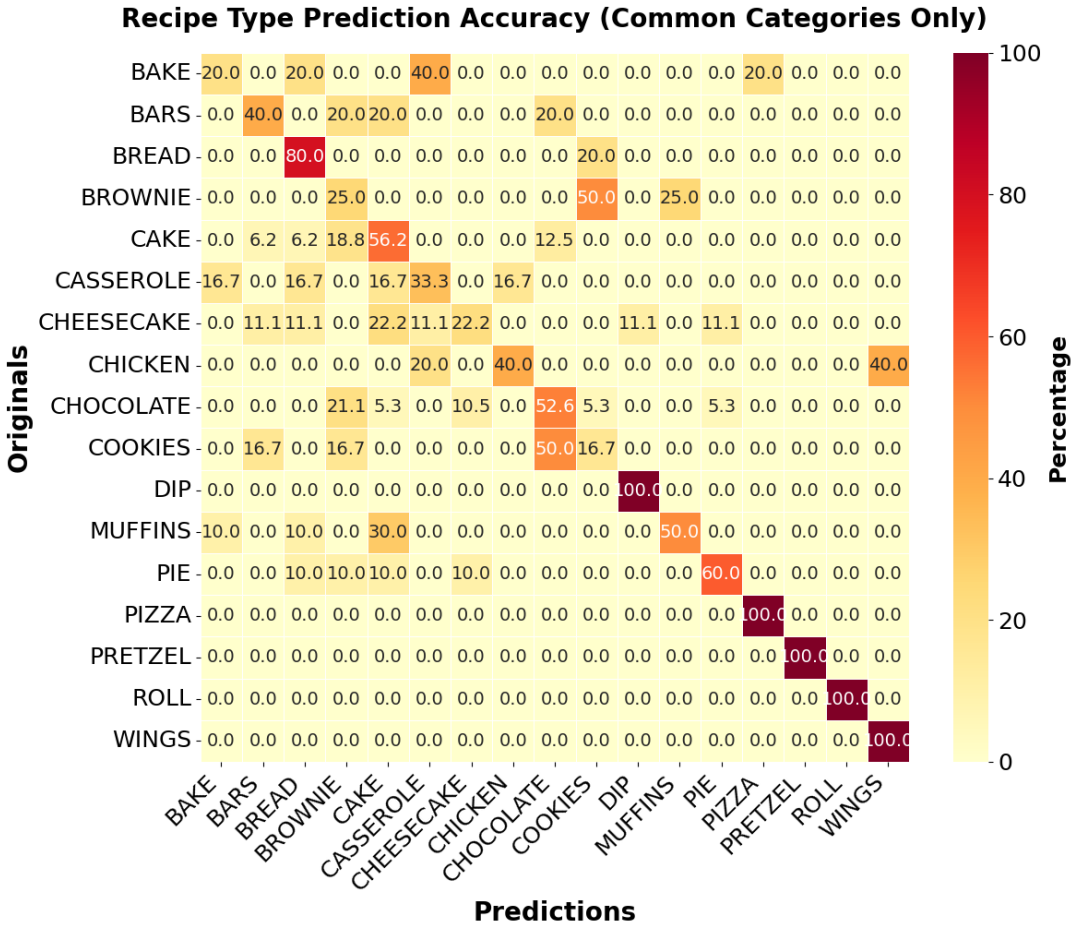
	it	fr	de	nl	sv	type - it	type - fr	type - de	type - nl	type - sv
inductiv	**0.71**	**0.67**	**0.83**	**0.34**	**0.45**	**0.8**	**0.76**	**0.9**	**0.53**	**0.67**
transductiv	0.24	0.14	0.54	0.31	-0.17	0.45	0.67	0.59	0.39	0.34

Results – Inductive GCN vs Transductive GCN

	Avg. Sim. Title INDUCTIVE	Avg. Sim. Title TRANSDUCTIVE	Processed	N/A	Multiple
fr	0.29	0.28	47	**1**	2
de	0.31	0.30	**49**	**1**	**0**
it	0.28	0.30	42	4	4
nl	**0.32**	**0.31**	45	3	2
sv	0.26	0.27	34	16	**0**
en	0.31	**0.31**	48	2	**0**

Results – Inductive GCN vs Transductive GCN





HUB ROMÂN DE INTELIGENȚĂ ARTIFICIALĂ

Conclusions

- The inductive method is proved to outperformed the transductive method on unseen data.
- The pipeline allows for the processing of multilingual data.
- Similarity formula needs to take into consideration the percentage of the common nodes.
- Ingredient overlapping is the most important addition in recipes matching.

References

[1] Y. W. C. C. J. W. X. W. Shirui Pan, Linhao Luo, “Unifying large language models and knowledge graphs: A roadmap,” arXiv preprint arXiv:2306.08302, 2023.[Online]. Available: https://arxiv.org/abs/2306.08302

[2] W. L. Hamilton, Graph Representation Learning, ser. Synthesis Lectures onArtificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2020.[Online]. Available: https://www.cs.mcgill.ca/˜wlh/grl book/files/GRL Book.pdf

[3] J. L. William L. Hamilton, Rex Ying, “Inductive representation learningon large graphs,” arXiv preprint arXiv:1706.02216, 2017. [Online]. Available: https://arxiv.org/pdf/1706.02216

HUB ROMÂN
DE INTELIGENȚĂ ARTIFICIALĂ

THANK YOU!