UNIVERSITY OF MANNHEIM
Data and Web Science Group

Evaluating Knowledge Generation and Self-Refinement Strategies for LLM-based Column Type Annotation

Keti Korini, Christian Bizer

ADBIS
The 29th European Conference on Advances in Databases and Information Systems 2025
Tampere, Finland

September 25, 2025

Column Type Annotation

- Sub-task of table interpretation which aims at understanding table semantics
- **Goal:** annotate the columns with the semantic type of the values contained in each column.
- **Use cases:** Important pre-processing step for data search and data integration in the context of data lakes.

RecipeName	RestrictedDiet	Duration
???	???	???
Asparagus and Arugula Salad	GlutenFreeDiet	PT30M
Dark/White Chocolate Ice Cream	VegetarianDiet	PT0H4M
Cheesy Baked Zucchini Fries	GlutenFreeDiet	PT30M

Existing CTA methods

UNIVERSITY OF MANNHEIM
Data and Web Science Group

1. link entities to knowledge graphs, e.g. DAGOBAH, MTab
2. fine-tune BERT or RoBERTa, e.g. TURL, DODUO
3. Prompt-based methods for LLMS, e.g. ArcheType, RACOON
4. Fine-tuning LLMs with focus on generalization, e.g. TableLlama, TableGPT, JellyFish

Liu et al..: **DAGOBAH: an end-to-end context-free tabular data semantic annotation system**. SemTab 2019.

Nguyen et al.: **Mtab: Matching tabular data to knowledge graph using probability models**. SemTab 2019.

Deng et al.: **TURL: Table understanding through representation learning**. VLDB 2020.

Suhara, et al.: **Annotating Columns with Pre-trained Language Models**. SIGMOD 2022.

Zhang, et al.: **TableLlama: Towards Open Large Generalist Models for Tables.** arXiv 2023.

Li et al.: **Table-GPT: Tabletuned GPT for Diverse Table Tasks**. arXiv 2023.

Feuer et al.: **ArcheType: A Novel Framework for OpenSource Column Type Annotation using Large Language Models**. arXiv 2023

Wei et al.: **Racoon: An llm-based framework for retrieval augmented column type annotation with a knowledge graph**. In: NeurIPS 2024 Third Table Representation Learning Workshop (2024)

Problem Statement

- The usage of labels slightly differs from dataset to dataset and domain-specific clues might be helpful for distinguishing between ambiguous labels
  - **General label:** “broadcast”
  - **Ambiguous labels:** “Review” and “Recipe Description”
- How to best adapt LLMs to how labels are used by specific datasets?

Our Contributions

UNIVERSITY OF MANNHEIM
Data and Web Science Group

1. we explore **knowledge generation prompting** for generating label definitions as a method to adapt the annotations to how terms are used by specific datasets.
2. we evaluate the **self-refining** abilities of the chosen LLMs by designing a pipeline to update the generated definitions based on errors made on a validation set
3. we evaluate the performance of **integrating the generated definitions** into the fine-tuning process
4. we compare fine-tuning and non-fine-tuning setups in terms of **efficiency** by comparing the **token usage** and by **performance** by comparing the **F1 score**.

Experimental Setup

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Datasets:
  - **SOTAB V2**: books, recipes, movies etc. (multi-class)
  - **WikiTURL**: film, broadcast, food, books etc. (multi-label)
  - **Limaye**: Wikipedia tables from books, people, etc. (multi-label)

	Train	Validation	Test	Columns	Avg. Columns	Labels
SOTAB V2	44K	456	609	1,851	2.62	82
SOTAB V2-ds	698	199	239	824	2.80	50
WikiTURL	397K	4,8K	4,7K	13K	1.60	255
WikiTURL-ds	809	416	379	878	1.44	66
Limaye	105	-	107	107	1	26

- LLMs tested:
  - **Llama models**: Llama-3.1-8B-4-bit and Llama-3.1-70B-4-bit
  - **OpenAI models**: gpt-4o-mini-2024-07-18 and gpt-4o-2024-03-15
  - **Fine-tuning models**: gpt-4o-2024-08-06

Baselines

1. Zero-shot Prompting

Task

Your task is to classify the columns of a given table with only one of the following classes that are separated with comma: [list of labels]

Instructions

Your instructions are: 1. Look at the cell values in detail. The first row of the table corresponds to the column names. 2. For each column, select one or more label/s that best represents the meaning of all cells in the column. The column can have multiple labels that have the same semantic meaning. 3. Answer with the selected label/s for each column using the JSON format {column_name: [label/s]}. 4. Answer only with labels from the provided label set!

Input

Classify these table columns: | Column 1 | Column 2 |
|:---|:---|
| " Achilles Last Stand " | Jimmy Page , Robert Plant |
| " All My Love " | John Paul Jones , Plant |

{ “Column 1”: ['written work’, 'music album’], “Column 2’: ['music artist’]}

Baselines

1. **Zero-shot Prompting**
2. **Few-shot Prompting**
   - Similarity-based demonstrations using *text-embedding-3-small*

TASK DESCRIPTION

TASK INSTRUCTIONS

Demonstration

Classify these table columns: | Column 1 | Column 2 | Column 3 |
Asparagus and Arugula Salad | GlutenFreeDiet | PT30M
Cheesy Baked Zucchini Fries | GlutenFreeDiet | PT30M

{ “Column 1”: “Recipe name”, “Column 3’: Duration}

INPUT TABLE

MODEL RESPONSE

Baselines

1. **Zero-shot Prompting**
2. **Few-shot Prompting**
3. **Self-consistency**
   - Self-consistency Prompting by Wang et al.
   - Run 3 times the same prompt for each input with **different temperatures.**
   - Take a **vote** on the 3 responses and get the final answer for each input.
   - In our paper, we use **temperatures 0, 0.5 and 0.7.**

Wang, X., Wei, J., Schuurmans, D., Le, Q.V., et al.: **Self-consistency improves chain of thought reasoning in language models.** In: The Eleventh International Conference on Learning Representations (2023)

Results: Baselines

Dataset	Setup	Micro-F1 Results				Inference Costs		
		Ll-8B	Ll-70B	gpt-mini	gpt-4o	Tokens	Cost	Cost/Col.
SOTAB V2	0-shot	56.0	67.4	69.4	80.9	270K	$0.87	$0.001
	self-con.	55.2	67.5	70.6	80.0	810K	$2.61	$0.004
	5-shot	59.3	64.6	63.4	81.8	1,047K	$2.73	$0.004
Limaye	0-shot	66.0	76.7	76.5	82.6	33K	$0.09	$0.0008
	self-con.	66.7	77.8	80.8	84.3	99K	$0.28	$0.002
	5-shot	90.2	81.5	91.2	89.4	90K	$0.23	$0.002
WikiTURL	0-shot	15.0	54.0	53.3	55.3	–	–	–
	5-shot	40.4	54.8	56.0	61.5	–	–	–
	0-hier	31.8	57.3	61.0	70.0	476K	$1.29	$0.001
	self-con.	36.3	58.4	61.6	69.2	1,428K	$3.89	$0.003
	5-hier	43.6	56.1	58.5	63.7	625K	$1.63	$0.002

- **Demonstrations** increase the performance only of the small **Llama-8B**, exception on Limaye.
- **Self-consistency** performs similar to zero-shot prompting but at triple the cost.

Knowledge Generation Prompting

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **Knowledge Prompting by Liu et al.:** Knowledge facts are generated and incorporated into the prompt

- **Our approach for CTA:** Generating label definitions and augmenting zero-shot prompts with them
- **Motivation:** Guide the generation of the model towards dataset-specific annotation

Liu, J., Liu, A., Lu, X., Welleck, S., et al.: **Generated knowledge prompting for commonsense reasoning**. arXiv preprint arXiv:2110.08387 (2021)

Knowledge Generation Prompting

- **Label Definitions:** One or more sentences describing the usage of the label.
  - Generated using GPT-4o (*gpt-4o-2024-03-15*).

- We experiment with **4 different types of definitions**
  1. Initial Definitions
  2. Demonstration-based Definitions
  3. Error-based Comparative Definitions
  4. Error-based Refined Definitions (Self-refinement Pipeline)

Knowledge Generation Prompting

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- We experiment with 4 different types of definitions
  1. **Initial Definitions**
     - Extract the knowledge that the LLM possesses about the labels
     - Baseline definitions, **not dataset-specific**

Your task is to generate definitions about some terms that can help in the task of column type annotation.

Generate a definition about the term ‘broadcast’.

The term "broadcast" refers to the distribution of audio or video content to a dispersed audience via any electronic mass communication medium, typically radio, television, or the internet. It involves transmitting signals over a specific frequency or through digital platforms to reach a wide audience simultaneously. Broadcasts can include **live events, pre-recorded programs, news, entertainment, and educational content**, and they play a crucial role in disseminating information and providing entertainment to the public.

Knowledge Generation Prompting

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- We experiment with 4 different types of definitions
  1. **Initial Definitions**

	Micro-F1 Results			
Method	**Llama-8B**	**Llama-70B**	**4o-Mini**	**GPT-4o**
SOTAB	55.0	64.3	71.7	79.3
Delta 0-shot	-1.0	-3.1	+1.7	-1.6
Limaye	63.6	77.8	75.7	82.8
Delta 0-shot	-2.4	+1.1	-0.8	+0.2
WikiTURL	29.8	47.5	60.4	69.8
Delta 0-shot	-2.0	-9.8	-0.6	-0.2

- **Result:** Overall, *initial* definitions have a negative impact on the model performance when compared to zero-shot prompting.

Knowledge Generation Prompting

- We experiment with 4 different types of definitions
  1. Initial Definitions
  2. **Demonstration-based Definitions**
     - Generated by showing model three demonstrations
     - **Dataset-specific** definition

Your task is to generate definitions about some terms that can help in the task of column type annotation.

Generate a definition about the term ‘broadcast’ using the following examples:

BookName | BookFormat | …
A Handbook for Morning Time | Paperback |…

A 'broadcast' refers to the transmission of audio or video content to a dispersed audience via any electronic mass communication medium, typically radio or television. **In the context of radio stations**, it denotes the specific radio station that is transmitting the content. **For television**, it can refer to the specific TV station or the network that is broadcasting the content. The term encompasses both the act of transmitting and the medium through which the content is delivered.

Knowledge Generation Prompting

- We experiment with 4 different types of definitions
  1. Initial Definitions
  2. **Demonstration-based Definitions**

	Micro-F1 Results			
Method	**Llama-8B**	**Llama-70B**	**4o-Mini**	**GPT-4o**
SOTAB	58.3	70.4	72.9	82.6
Delta 5-shot	-1.0	+5.8	+9.5	+0.8
Limaye	66.1	81.6	77.7	86.8
Delta 5-shot	-24.1	+0.1	-13.5	-2.6
WikiTURL	35.3	51.5	61.6	72.2
Delta 5-shot	-8.3	-4.6	+3.1	+8.5

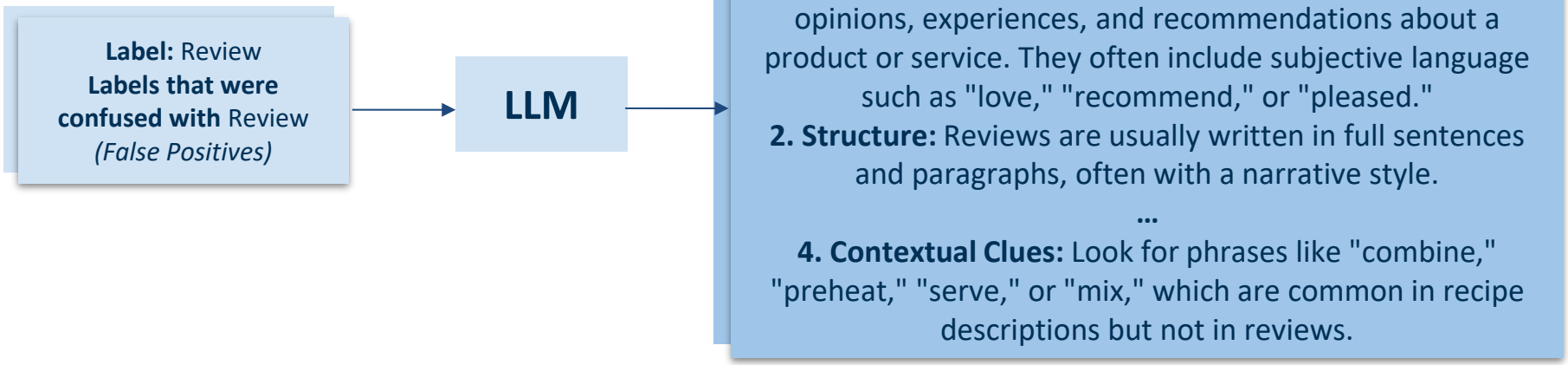
- **Result:** For larger models, F1 increases compared to using the demonstrations directly in few-shot prompting.

Knowledge Generation Prompting

- We experiment with 4 different types of definitions
  1. Initial Definitions
  2. Demonstration-based Definitions
  3. **Error-based Comparative Definitions**
     - Motivation: Provide comparisons of labels that the LLM uses wrongly
     - Pair-wise definitions/comparisons

1. Classification of validation set with zero-shot prompt

2. Generate comparative definitions



Knowledge Generation Prompting

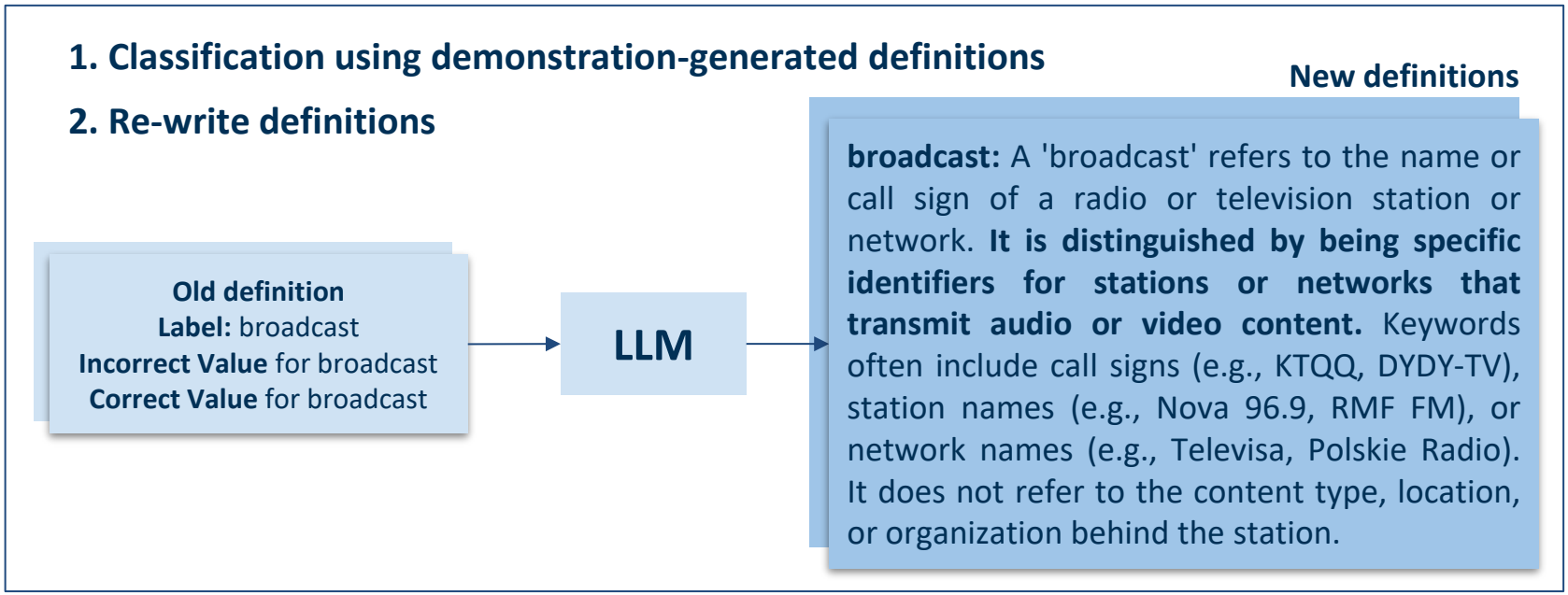
- We experiment with 4 different types of definitions
  1. Initial Definitions
  2. Demonstration-based Definitions
  3. **Error-based Comparative Definitions**

	Micro-F1 Results			
Method	**Llama-8B**	**Llama-70B**	**4o-Mini**	**GPT-4o**
SOTAB	55.4	64.1	70.3	83.7
Delta 5-shot	-3.9	-0.5	+6.9	+1.9
Limaye	77.3	76.6	80.5	85.0
Delta 5-shot	-12.9	-4.9	-10.7	-4.4
WikiTURL	33.2	48.8	62.2	71.6
Delta 5-shot	-10.4	-7.3	+3.7	+7.9

- **Result:** For the smaller models performance decreases, while for the larger models we have increases in 2 out of 3 datasets.

Knowledge Generation Prompting

- We experiment with 4 different types of definitions
  1. Initial Definitions
  2. Demonstration-based Definitions
  3. Error-based Comparative Definitions
  4. **Error-based Refined Definitions (Self-refinement Pipeline)**
     - Labels are updated based on errors made in the validation set when using the demonstration-based definitions



Knowledge Generation Prompting

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- We experiment with 4 different types of definitions
  1. Initial Definitions
  2. Demonstration-based Definitions
  3. Error-based Comparative Definitions
  4. **Error-based Refined Definitions (Self-refinement Pipeline)**

	Micro-F1 Results			
Method	**Llama-8B**	**Llama-70B**	**4o-Mini**	**GPT-4o**
SOTAB	59.8	72.0	75.1	85.4
Delta 5-shot	+0.5	+7.4	+11.7	+3.6
Limaye	79.1	84.9	80.2	88.4
Delta 5-shot	-11.1	+3.4	-10.7	-1.0
WikiTURL	42.6	49.3	65.5	72.4
Delta 5-shot	-1.0	-4.6	+7.0	+8.6

- **Result:** For the smaller models performance decreases, while for the larger models we have increases in 2 out of 3 datasets.

Knowledge Generation Prompting Results Summary

- Using **demonstrations** in the prompt or using demonstrations to generate **definitions**?
  - For OpenAI models, 0.8-9.5% increase in two datasets out of three when using definitions.
- Was the **self-refinement pipeline** effective?
  - In all cases except for 2, there is an average increase of 3.9% to the F1 score when refining the *demonstration* definitions.

- Overall, *refined* definitions have the highest performance among the different types of definitions tested.

Fine-tuning Setups

- We evaluate three fine-tuning setups:
  - Simple fine-tuning
  - Multi-task fine-tuning: combine CTA task with knowledge generation
  - Multi-task fine-tuning with demonstrations: similar to above, but in the knowledge generation prompt add 3 demonstrations
- We test the fine-tuned models using **zero-shot prompting** and **knowledge prompting.**

Results: Fine-tuning

Setup	SOTAB V2				WikiTURL			
	Ll-8B	Ll-70B	mini	gpt-4o	Ll-8B	Ll-70B	mini	gpt-4o
Zero-shot Prompting								
simple fine-tuning	77.9	86.4	87.0	87.8	62.5	64.3	**67.3**	71.1
multi-task fine-tuning	77.3	**87.8**	87.0	-	**64.1**	**66.4**	64.4	-
multi-task-with-demos	**80.7**	86.9	-	-	63.1	60.2	-	-
Δ 0-shot (not fine-tuned)	+24.7	+20.4	+17.6	+6.9	+32.3	+9.1	+6.3	+1.1
Prompting with Definitions								
FT + comparative defs	80.8	85.8	**89.1**	90.0	**70.1**	**65.6**	65.4	**74.1**
FT + demonstration defs	80.3	85.6	87.0	89.1	65.0	65.0	**66.4**	72.3
FT + refined defs	**81.3**	**86.1**	87.2	**91.8**	65.1	64.4	63.3	74.0
Δ 0-shot fine-tuned	+0.6	-1.7	+2.1	+4.0	+6.0	-0.8	-0.9	+3.0

- Fine-tuning with combination of CTA and knowledge generation benefits the Llama models in small percentages.
- **GPT-4o benefits from knowledge prompting** on both datasets with an increase of at least 3% in F1 score.
- Both *comparative* and *refined* definitions bring this increase in F1, however the ***refined* definitions** are **better token-wise**.

Fine-tuning or Knowledge Prompting?

UNIVERSITY OF MANNHEIM
Data and Web Science Group

		F1	Costs			
Method	**Setup**	**GPT-4o**	**FT Cost**	**Generation Cost**	**Inference Cost**	**Cost/Column**
SOTAB	Refined	85.4	-	$3.50	$4.27	$0.007
	Ft-0-shot	87.8	$47.4	-	$1.28	$0.002
WikiTURL	Refined	72.4	-	$8.36	$12.1	$0.016
	Ft-0-shot	71.1	$20.0	-	$1.88	$0.002

Fine-tuning or knowledge prompting without fine-tuning?

- Total cost of knowledge prompting is lower than fine-tuning *for the chosen datasets*.
- However, **fine-tuning** becomes more cost efficient in cases with larger amount of tables as the **inference cost is lower** e.g. in cases with more than 9400 columns to be annotated (using SOTAB as reference) .

Zero or Knowledge Prompting in Fine-tuning Setup?

		F1	Costs			
Method	**Setup**	**GPT-4o**	**FT Cost**	**Generation Cost**	**Inference Cost**	**Cost/Column**
SOTAB	Ft-0-shot	87.8	$47.4	-	$1.28	$0.002
	Ft-refined	91.8	$47.4	$3.48	$3.90	$0.007
WikiTURL	Ft-0-shot	71.1	$20.0	-	$14.3	$0.002
	Ft-refined	74.0	$20.0	$10.9	$16.8	$0.020

Zero-shot or knowledge prompting when using fine-tuned models?

- **GPT-4o:** Knowledge prompting brings at least 3% increase in F1 on both datasets tested.
- **Smaller models:** Similar F1 score to zero-shot prompting.

Conclusions

- We tested two methods for adapting the LLM generation to the datasets used for testing: **knowledge generation prompting** and a **self-refinement** pipeline.
- The **generated definitions** increase the F1 score in most cases by an average of 2.4% compared to zero-shot prompting.
- Further **refining these definitions** brings an additional average increase of 3.9% in most cases.
- We conclude that fine-tuning is more token efficient for use cases with large number of tables than using refined definitions.
- Fine-tuned GPT-4o benefits of an additional 3% increase when combined with knowledge generation prompting.

Thank you.

UNIVERSITY OF MANNHEIM
Data and Web Science Group

GitHub link:

https://github.com/wbsg-uni-mannheim/TabAnnGPT