Automating Benchmarking Process for Multimodal Large Language Models (MLLMs) in the Context of Waste Disposal

Sundus Hammoud
Technical University of Clausthal
Institute for Software and Systems Engineering
sundus.hammoud@tu-clausthal.de

Robert Werner
Technical University of Clausthal
Institute for Software and Systems Engineering
robert.werner@tu-clausthal.de

02.10.2024

Hello!

I am Sundus Hammoud

- Graduated from Damascus University in Syria, department of Information Technology.
- Master student in the Institute of software and systems technology in Technical University of Clausthal in Germany.
- Software developer in Ceconsoft.

Content

- Motivation
- Possible solution and challenges
- Proposed approach
- Dataset collection
- Prompts in LLMs
- LLMs evaluating LLMs
- Results
- Future outlook
- Conclusion

Motivation

- 60% of German citizens lack detailed information on the correct disposal and therefore throw garbage in the wrong bins.

- People don’t know who to ask to gain knowledge on which bin to use.

Motivation

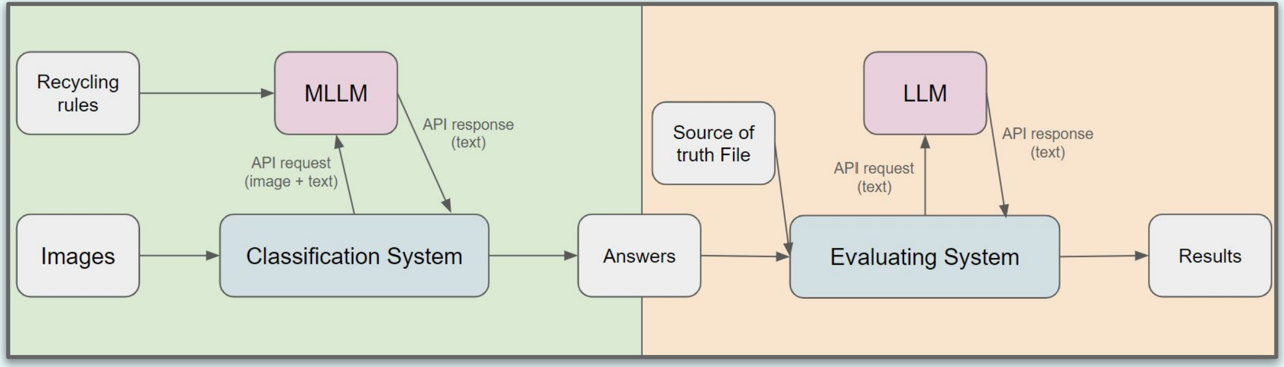
- What happens when people don’t recycle properly?
    - This hinders circular economy because we will exploit non-renewable resources to produce materials like plastic.
    - Mixing different materials can contaminate the whole batch, making it unusable.
    - Growing waste landfill sizes and therefore polluting the environment with the emissions of toxic greenhouse gases.

Challenges

- How can we tackle this situation?
  - Introducing chatbots to interact with users where they can ask recycling questions.
  - Garbage scanning feature that enables the user to upload an image rather than explaining in text.
- Each city in Germany have separate rules for recycling, which means we cannot rely on the general knowledge that the model could be trained on. For example: organic waste
- LLMs respond in long and detailed chunks for text, this is very hard for a human being to go through thousands of text paragraphs and see if the answer meets the expectations.
- The development process is iterative, enhancing the algorithm of tweaking the models parameters require repeating the process numerous iterations.

Proposed approach

- Automate the benchmarking process in the context of waste recycling.
- This problem will be tackled from two different sides:
  - Resolving the region-based recycling rules.
  - Minimizing the human role by introducing non-human judge.



Benchmarking system architecture

Dataset Collection

- There are 6 categories represented in this dataset, which are yellow, blue, green and black. clothes containers and glass containers.

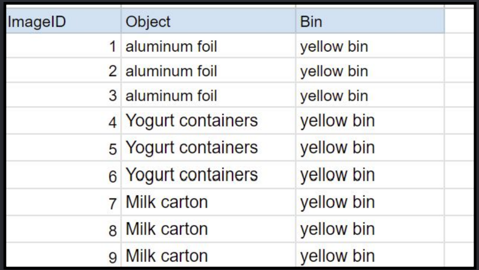
- Wolfenbüttel has a document on its recycling rules and a website where all information are up-to-date.

- The images should contain objects that are in a realistic environment, so no white background where an item is pretty isolated.

Source: google images + kleinanzeigen website
https://www.kleinanzeigen.de/

Dataset Collection

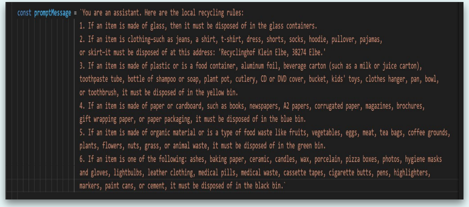
- For each object inside the list, 3 photos are collected manually using Google’s image search or Kleinanzeigen website.
- There are a total of 207 pictures that span the 6 mentioned categories.
- We follow the rules in the official document of recycling released by Wolfenbüttel to define the source of truth file.



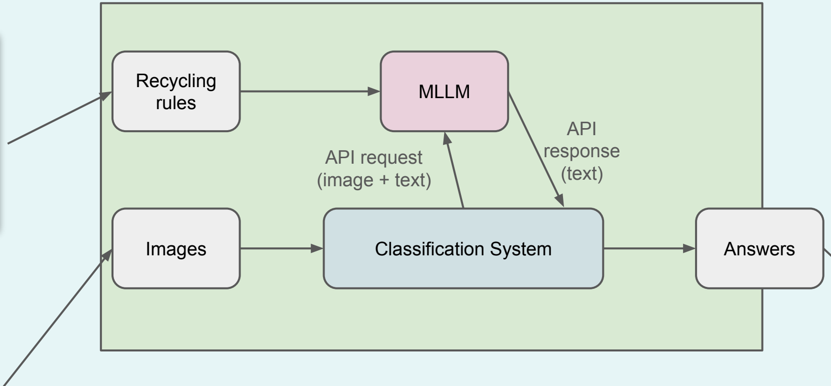
ImageID	Object	Bin
1	aluminum foil	yellow bin
2	aluminum foil	yellow bin
3	aluminum foil	yellow bin
4	Yogurt containers	yellow bin
5	Yogurt containers	yellow bin
6	Yogurt containers	yellow bin
7	Milk carton	yellow bin
8	Milk carton	yellow bin
9	Milk carton	yellow bin

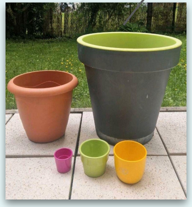
A snippet of the source of truth file

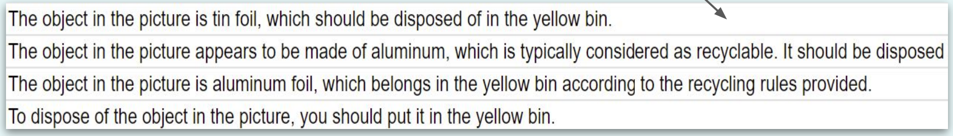
Classification phase



```
const promptMessage = `You are an assistant. Here are the local recycling rules:
                      1. If an item is made of glass, then it must be disposed of in the glass containers.
                      2. If an item is clothing-such as jeans, a shirt, t-shirt, dress, shorts, socks, hoodie, pullover, pajamas,
                      or skirt-it must be disposed of at this address: 'Recyclinghof Klein Elbe, 38274 Elbe.'
                      3. If an item is made of plastic or is a food container, aluminum foil, beverage carton (such as a milk or juice carton),
                      toothpaste tube, bottle of shampoo or soap, plant pot, cutlery, CD or DVD cover, bucket, kids' toys, clothes hanger, pan, bowl,
                      or toothbrush, it must be disposed of in the yellow bin.
                      4. If an item is made of paper or cardboard, such as books, newspapers, A2 papers, corrugated paper, magazines, brochures,
                      gift wrapping paper, or paper packaging, it must be disposed of in the blue bin.
                      5. If an item is made of organic material or is a type of food waste like fruits, vegetables, eggs, meat, tea bags, coffee grounds,
                      plants, flowers, nuts, grass, or animal waste, it must be disposed of in the green bin.
                      6. If an item is one of the following: ashes, baking paper, ceramic, candles, wax, porcelain, pizza boxes, photos, hygiene masks
                      and gloves, lightbulbs, leather clothing, medical pills, medical waste, cassette tapes, cigarette butts, pens, highlighters,
                      markers, paint cans, or cement, it must be disposed of in the black bin.`
```

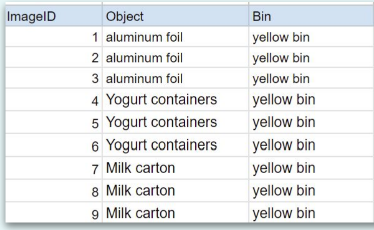




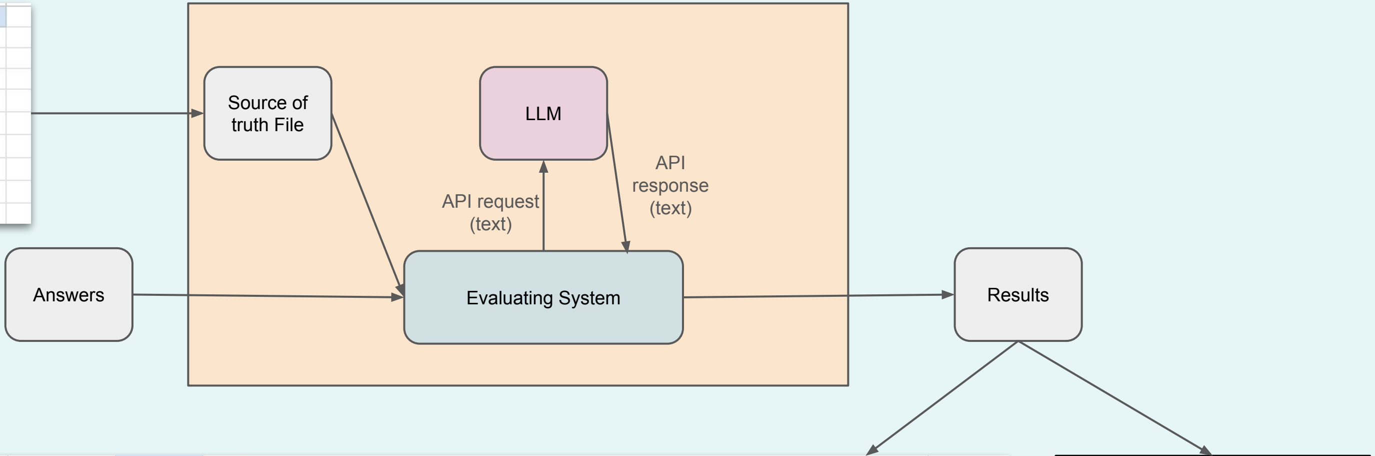


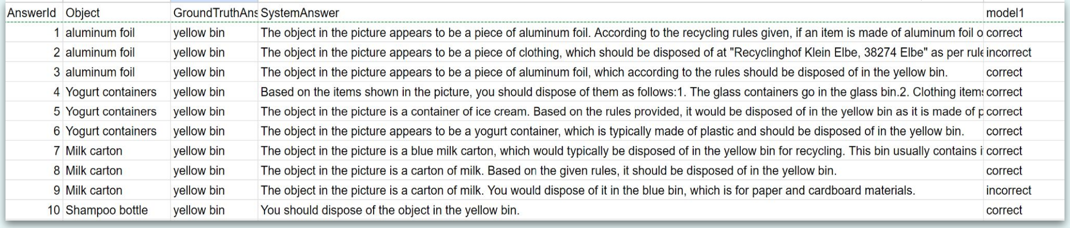
The object in the picture is tin foil, which should be disposed of in the yellow bin.
The object in the picture appears to be made of aluminum, which is typically considered as recyclable. It should be disposed
The object in the picture is aluminum foil, which belongs in the yellow bin according to the recycling rules provided.
To dispose of the object in the picture, you should put it in the yellow bin.

Evaluation phase



ImageID	Object	Bin
1	aluminum foil	yellow bin
2	aluminum foil	yellow bin
3	aluminum foil	yellow bin
4	Yogurt containers	yellow bin
5	Yogurt containers	yellow bin
6	Yogurt containers	yellow bin
7	Milk carton	yellow bin
8	Milk carton	yellow bin
9	Milk carton	yellow bin





AnswerId	Object	GroundTruthAns	SystemAnswer	model1
1	aluminum foil	yellow bin	The object in the picture appears to be a piece of aluminum foil. According to the recycling rules given, if an item is made of aluminum foil o	correct
2	aluminum foil	yellow bin	The object in the picture appears to be a piece of clothing, which should be disposed of at "Recyclinghof Klein Elbe, 38274 Elbe" as per rule	incorrect
3	aluminum foil	yellow bin	The object in the picture appears to be a piece of aluminum foil, which according to the rules should be disposed of in the yellow bin.	correct
4	Yogurt containers	yellow bin	Based on the items shown in the picture, you should dispose of them as follows:1. The glass containers go in the glass bin.2. Clothing item	correct
5	Yogurt containers	yellow bin	The object in the picture is a container of ice cream. Based on the rules provided, it would be disposed of in the yellow bin as it is made of p	correct
6	Yogurt containers	yellow bin	The object in the picture appears to be a yogurt container, which is typically made of plastic and should be disposed of in the yellow bin.	correct
7	Milk carton	yellow bin	The object in the picture is a blue milk carton, which would typically be disposed of in the yellow bin for recycling. This bin usually contains i	correct
8	Milk carton	yellow bin	The object in the picture is a carton of milk. Based on the given rules, it should be disposed of in the yellow bin.	correct
9	Milk carton	yellow bin	The object in the picture is a carton of milk. You would dispose of it in the blue bin, which is for paper and cardboard materials.	incorrect
10	Shampoo bottle	yellow bin	You should dispose of the object in the yellow bin.	correct

Incorrect answers count

Prompts in LLM

- A prompt is a text through which a human being can interact with the language model, its purpose is to give instructions or context information to the language model.

- There are two types of prompts:
  System Prompt: is a prompt that influences the entire model's behavior, it could be a set of rules to follow or some information related to a context. This prompt is given to the model only once before any user interaction.

  User Prompt: is when a prompt is given to the system while expecting an answer, through which the user usually interacts with the model.

- The system prompt a very good solution for the regional recycling rules problem, as we can feed the model this context knowledge.

Prompts in LLM

- The prompting strategy that is used for both models is called “Persona”.
- It gives the model a personality with perspective and knowledge on how to act if a user ask it a question.
- In the prompt construction we notice the following elements:
  - The role the model will play.
  - the context knowledge defined as a set of rules.
  - Examples for the model to reason about.

“You are an assistant. Here are the local recycling rules:

1. If an item is made of glass, then it must be disposed of in the glass containers.

2. If an item is clothing--such as jeans, a shirt, t-shirt, dress, shorts, socks, hoodie, pullover, pajamas, or skirt--it

must be disposed of at this address: ’Recyclinghof Klein Elbe, 38274 Elbe.’

3. If an item is made of plastic or is a food container, aluminum foil, beverage carton (such as a milk or juice carton), toothpaste tube, bottle of shampoo or soap, plant pot, cutlery, CD or DVD cover, bucket, kids’ toys, clothes hanger, pan, bowl, or toothbrush, it must be disposed of in the yellow bin.

…”

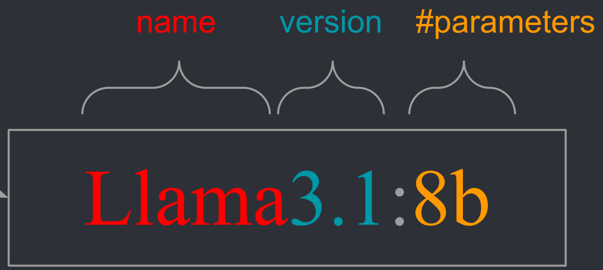
LLMs evaluating LLMs

- We can use the prompting capabilities of an LLM to turn it into a judge to evaluate the first LLM’s answers.
- The judge model can’t rely on the general knowledge that it was trained with, because the classification system is making decisions regarding the information that was fed into it using the prompt.
- For the prompt: a combination of “Persona” and “Reference-guided grading” strategies.
- This approach means the system is provided with a reference answer, and another system’s answer, and the model must compare if they are semantically equal.

> *"You are an evaluation assistant. Your task is to compare two texts: the first text contains the source of truth, and the second text contains system answers. Determine if the bin mentioned in the source of truth matches the bin mentioned in the system answer. Respond with one word: 'correct' if they match, and 'incorrect' if they don't."*

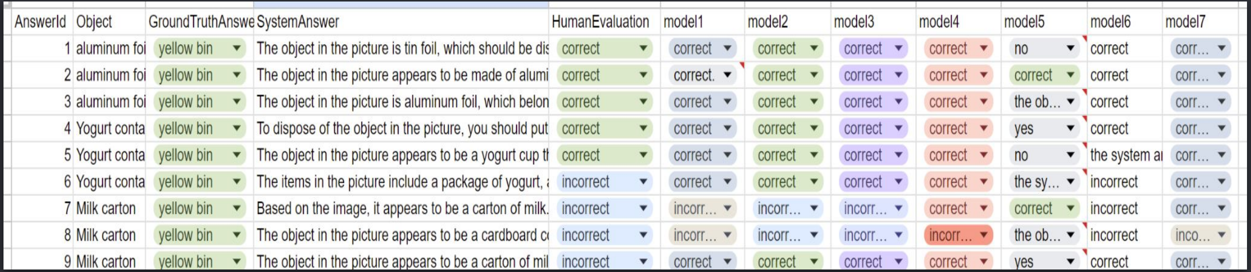
Who is our judge model?

- The benchmarking system will be as a good as its judge.
- We need to know which model is best judge.
- The following models claim to have high reasoning capabilities:
  - Llama3.1:8b
  - Qwen2:7b
  - Llama3: 8b
  - Qwen:7b
  - Mistrallite:7b
  - Llama3.1:70b
  - Qwen:32b



Who is our judge model?

- LLMs produce non-deterministic responses, this means that we may receive different results when we interact with them.
- we run the classification phase once to obtain the answers and then label them by a human
- we run the evaluation models three times.
- We compute the confusion matrix and then the accuracy, precision, recall and f1 measure for each model.

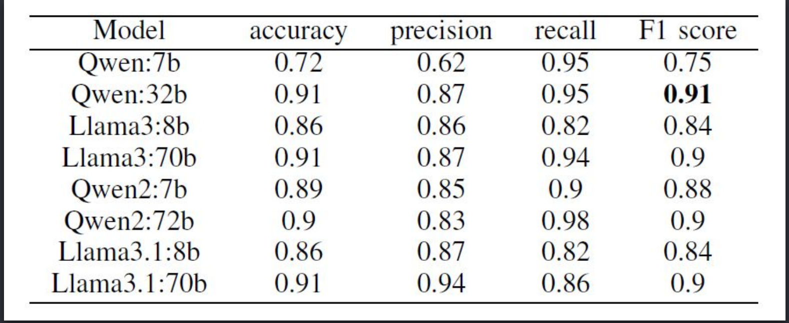


AnswerId	Object	GroundTruthAnswe	SystemAnswer	HumanEvaluation	model1	model2	model3	model4	model5	model6	model7
1	aluminum foi	yellow bin	The object in the picture is tin foil, which should be dis	correct	correct	correct	correct	correct	no	correct	corr...
2	aluminum foi	yellow bin	The object in the picture appears to be made of alumi	correct	correct.	correct	correct	correct	correct	correct	corr...
3	aluminum foi	yellow bin	The object in the picture is aluminum foil, which belon	correct	correct	correct	correct	correct	the ob...	correct	corr...
4	Yogurt conta	yellow bin	To dispose of the object in the picture, you should put	correct	correct	correct	correct	correct	yes	correct	corr...
5	Yogurt conta	yellow bin	The object in the picture appears to be a yogurt cup tl	correct	correct	correct	correct	correct	no	the system ai	corr...
6	Yogurt conta	yellow bin	The items in the picture include a package of yogurt, ;	incorrect	correct	correct	correct	correct	the sy...	incorrect	corr...
7	Milk carton	yellow bin	Based on the image, it appears to be a carton of milk.	incorrect	incorr...	incorr...	incorr...	correct	correct	incorrect	corr...
8	Milk carton	yellow bin	The object in the picture appears to be a cardboard c	incorrect	incorr...	incorr...	incorr...	incorr...	the ob...	incorrect	inco...
9	Milk carton	yellow bin	The object in the picture appears to be a carton of mil	incorrect	correct	correct	correct	correct	yes	correct	corr...

A snippet of the assembled output of all the models in addition to the classification system's answers and the human evaluation.

Results

- We see that the model Qwen:32b outperform all the others.
- We notice in Llama and Qwen model families, higher parameters count means a better performance.
- A high number of parameters may ensure a better performance but the model is slightly slower.
- The incorrect answers count using the Qwen:32b model is 109/207.
- While the incorrect answers annotated by a human 116/207.
- Qwen:32b gives a good estimation of the situation.



Model	accuracy	precision	recall	F1 score
Qwen:7b	0.72	0.62	0.95	0.75
Qwen:32b	0.91	0.87	0.95	**0.91**
Llama3:8b	0.86	0.86	0.82	0.84
Llama3:70b	0.91	0.87	0.94	0.9
Qwen2:7b	0.89	0.85	0.9	0.88
Qwen2:72b	0.9	0.83	0.98	0.9
Llama3.1:8b	0.86	0.87	0.82	0.84
Llama3.1:70b	0.91	0.94	0.86	0.9

Results table

Future work

- **Keeping up-to-date:** these model are always under maintenance and new models are always under development.
- **Prompt Engineering:** There are a set of techniques for structuring the best prompt, it would boost the performance once this topic can be applied with more depth.
- **Categories-oriented evaluation:** output incorrect count for the individual categories can give better understanding on where the model is failing.
- **Separate models for separate tasks:** In the classification phase, we can use a models that is better in recognising object, and another model which better in reasoning about rules and prompts.
- **Benchmarking framework:** with an interface for researchers, they can upload the ground of truth file and dataset, edit the prompt, select classification models, run different models and compare their results.

Conclusion

- While some may remain cautious about fully trusting AI-driven evaluations, many are becoming increasingly open to leveraging large language models.

- Ceconsoft is currently implementing a project where an MLLM is needed and a benchmarking approach is needed, so we will utilize this benchmarking approach to verify the final product is ready for use.

- Would you as researchers feel confident relying on these tools to benchmark your algorithms, or would you still prefer to evaluate your work manually?

Thank you!

You can always reach us at:
sundus.hammoud@tu-clausthal.de
robert.werner@tu-clausthal.de