

Table Evaluation processing

Pablo de Vicente. 10/12/24

This document evaluates various LLMs for generating descriptions of table data. The tables tested vary in type to reflect real-world scenarios. To ensure consistency, identical prompts are used for all tables. Each table is provided in its entirety in CSV format rather than as individual rows, preserving context and improving processing efficiency.

prompt used: "Describe in a technical manner each of the rows of the table"

Table 1:

Environmental Suitability:	Indoor - Riser, Indoor
:	:
Flammability / Reaction to Fire:	UL 1666 Riser, FT4, IEC 60332-1-2
CPR Compliance:	CPR Euroclass: Eca; CPR UKCA Class: Eca
NEC / UL Compliance:	Article 800, CMR
CEC / C(UL) Compliance:	CMG
European Directive Compliance:	EU CE Mark, EU Directive 2015/863/EU (RoHS 2 amendment), EU Directive 2011/65/EU (RoHS 2), EU Directive 2012/19/EU (WEEE)
UK Regulation Compliance:	UKCA Mark
APAC Compliance:	China RoHS II (GB/T 26572-2011)
Plenum Number:	6300UE

Using Llama 3.2:

3.5m x 3.5m x 1.5m
Lighting type: 3 x 60W CFLs
Floor type: Carpet

Ceiling type: Polished plaster

Wall type: Polished plaster

The flammability rating of this product is 1. The flammability rating is based on the flammability of the material when exposed to an open flame. The flammability rating is determined by the flammability of the material

1. A 100% recycled polyethylene pipe is used for the pipe. 2. The pipe is a 3/4 inch (19.05 mm) diameter pipe. 3. The pipe is a 3/4 inch (

1. A single-phase, 3-wire, 120/240 V, 50/60 Hz, 4.5 A, 1.5 kW, 60 Hz, 3-phase, 60 Hz, 120/

The following description was generated by the

The European Union (EU) has set a number of environmental requirements for products, which include the requirement to be RoHS-compliant and WEEE-compliant. The RoHS directive (2002/95/EC) applies to all electrical and

1. The location of the system is in a building that is located in an area that is subject to a high level of ambient noise. 2. The system is located in a building that is subject to a high level of ambient noise.

1. The description is generated from the

6300UE is a 12" x 12" x 12" (300 mm x 300 mm x 300 mm) plastic enclosure with a 6300U model number and a height of 12" (300 mm). The

Using bloom (pretrained on toto dataset):

The input format for this model required a structure closer to HTML or Markdown rather than CSV. The responses are also generated in this format. While this is not an issue since we can transform it later, no transformation was performed here as the model's inability to understand the data is already evident. The TOTO dataset primarily consists of Wikipedia tables, featuring information about famous people and other data types unrelated to our samples. Unfortunately, no public dataset of tables containing industrial products (or similar) was available for pretraining BLOOM.

```
'<s> <table> <row> <c> Environmental Suitability: </c> <c> Flammability / Reaction to Fire:
</c> <c> CPR Compliance: </c> <c> NEC / UL Compliance: </c> <c> CEC / C(UL) Compliance: </c>
<c> European Directive Compliance: </c> <c> UK Regulation Compliance: </c> <c> APAC
Compliance: </c> </row> <row> <c> Indoor - Riser, Indoor <col_header> Environmental
Suitability: </col_header> </c> <c> UL 1666 Riser, FT4, IEC 60332-1-2 <col_header>
Flammability / Reaction to Fire: </col_header> </c> <c> CPR Euroclass: Eca; CPR UKCA Class:
Eca <col_header> CPR Compliance: </col_header> </c> <c> Article 800, CMR <col_header> NEC / UL
Compliance: </col_header> </c> <c> CMG <col_header> CEC / C(UL) Compliance: </col_header> </c>
<c> EU CE Mark, EU Directive 2015/863/EU (RoHS 2 amendment), EU Directive 2011/65/EU (RoHS 2),
EU Directive 2012/19/EU (WEEE) <col_header> European Directive Compliance: </col_header> </c>
<c> UKCA Mark <col_header> UK Regulation Compliance: </col_header> </c> <c> China RoHS II
(GB/T 26572-2011) <col_header> APAC Compliance: </col_header> </c> </row> </table> </s>'Neon,
the Beast of Haarlem </page_title> <section_title> Awards and nominations </section_title>
<table> <row> <c> Award </c> <c> Category <row_header> Award </row_header> </c> <c> Recipient
<row_header> Award </row_header> <row_header> Category </row_header> </c> <c> Result
<row_header> Award </row_header> <row_header> Category </row_header> <row_header> Recipient
</row_header> </c> </row> <row> <c> Academy Award <col_header> Award </col_header> </c> <c>
Best Actor <col_header> Category </col_header> </c> <c> Peter Finch <col_header> Recipient
</col_header> </c> <c> Nominated <col_header> Result </col_header> </c> </row> <row> <c> Best
Actress <col_header> Award </col_header> </c> <c> Michelle Williams <col_header> Category
</col_header> </c> <c> Nominated <col_header> Recipient </col_header> </c> </row> <row> <c>
Best Supporting Actor <col_header> Award </col_header> </c> <c> David Fincher <col_header>
Category </col_header> </c> <c> Nominated <col_header> Recipient </col_header> </c> </row>
<row> <c> Best Supporting Actress <col_header> Award </col_header> </c> <c> Michelle Williams
<col_header> Category </col_header> </c> <c> Nominated <col_header> Recipient </col_header>
</c> </row> <row> <c> Best Supporting Actress <col_header> Award </col_header> </c> <c> Maggie
Smith <col_header> Category </col_header> </c> <c> Nominated <col_header> Recipient
```

	Best Supporting Actor	Award
David Fincher	Category	Nominated
		Recipient
	Best Supporting Actress	Award
Maggie Smith	Category	Nominated
		Recipient
	Best Supporting Actress	Award
Michelle Williams	Category	Nominated
		Recipient
	Best Supporting Actress	Award
Maggie Smith	Category	Nominated
		Recipient
	Best Supporting Actor	Award
David Fincher	Category	Nominated
		Recipient
	Best Supporting Actress	Award
Michelle Williams	Category	Nominated
		Recipient
	Best Supporting Actress	Award
Maggie Smith	Category	Nominated
		Recipient
	Best Supporting Actress	Award
Michelle Williams	Category	Nominated
		Recipient
	Best Supporting Actress	Award
Maggie Smith	Category	Nominated
		Recipient
	Best Supporting Actor	Award
David Fincher	Category	Nominated
		Recipient
	Best Supporting Actress	Award
Michelle Williams	Category	Nominated
		Recipient
	Best Supporting Actress	Award
Maggie Smith	Category	Nominated
		Recipient
	Best Supporting Actress	Award
Michelle Williams	Category	Nominated
		Recipient
	Best Supporting Actress	Award
Maggie Smith	Category	Nominated
		Recipient
	Best Supporting Actress	Award
Michelle Williams	Category	Nominated
		Recipient
	Best Supporting Actress	Award
Maggie Smith	Category	Nominated
		Recipient
	Best Supporting Actress	Award
Maggie Smith	Category	Nominated
		Recipient

Table 2:

Item #	Color	Putup Type	Length	UPC/EAN	
5300UE 0101000	Black	Reel	1,000 ft	612825159162	C
5300UE 010U1000	Black	UnReel	1,000 ft	612825421511	
5300UE 010Z1000	Black	ReelTuff Box	1,000 ft	612825159254	
5300UE 0061000	Blue, Light	Reel	1,000 ft	612825350644	
5300UE 001U1000	Brown	UnReel	1,000 ft	612825421450	
5300UE 001Z1000	Brown	ReelTuff Box	1,000 ft	612825350682	
5300UE.00152	Chrome	Reel	152 m	8719605008176	
5300UE.00305	Chrome	Reel	305 m	8719605008183	
5300UE 008500	Gray	Reel	500 ft	612825159117	
5300UE 008C500	Gray	Coil	500 ft	612825336761	
5300UE 008U500	Gray	UnReel	500 ft	612825421498	
5300UE 0081000	Gray	Reel	1,000 ft	612825159148	C
5300UE 008U1000	Gray	UnReel	1,000 ft	612825159124	
5300UE 008Z1000	Gray	ReelTuff Box	1,000 ft	612825159155	
5300UE 005Z1000	Green	ReelTuff Box	1,000 ft	612825350637	
5300UE 0051000	Green, Dark	Reel	1,000 ft	612825421474	
5300UE 0021000	Red	Reel	1,000 ft	612825421467	
5300UE 002Z1000	Red	ReelTuff Box	1,000 ft	612825350552	
5300UE 0071000	Violet	Reel	1,000 ft	612825421481	
5300UE 007Z1000	Violet	ReelTuff Box	1,000 ft	612825159131	
5300UE 0091000	White	Reel	1,000 ft	612825336778	C
5300UE 009U1000	White	UnReel	1,000 ft	612825421504	
5300UE 009Z1000	White	ReelTuff Box	1,000 ft	612825159179	

```
| 5300UE 0041000 | Yellow | Reel | 1,000 ft | 612825350590 |  
|
```

Llama 3.1: The model's inability to understand the data is evident, as it fabricates information and introduces inaccuracies along the way. The output is not altered in structure.

5 minutes ago, a 32 year old male patient presented to the emergency room with complaints of severe abdominal pain and nausea. He had a history of alcoholism and was taking 4 pills of alcohol a day. On physical exam, the patient was

1,000 ft of 0.045" diameter 1000 ft of 0.045" diameter 1000 ft of 0.045" diameter 1000 ft of 0.045" diameter 1000 ft of

1,000 ft of 10/0 UnReel Nylon Fishing Line

Item #: 5300UE 010U1000, Color: Black, Putup Type: UnReel, Length: 1,000 ft, UPC/E

1,000 ft. of 0.030" diameter, 0.080" thick, 100% virgin polyethylene (PE) tubing. ReelTuff boxes are designed to hold reels of 1/2" diameter tubing

1,000 ft. Blue, Light Putup. Reel of 1000 ft. Blue, Light, Putup. Each putup includes 100 ft. of 1/4" wire. 1/4"

1,000 ft of 0.5mm un-reel yarn in brown.

1,000 ft. of 30 lb. polyester rope

Dimensions: 10 ft. x 10 ft. x 0.125 in.

Item #: 5300UE 001Z1000, Color: Brown, Putup Type

1.0 m reel of 152 m (500 ft) of 0.1 mm (0.004 in.) diameter 304 stainless steel wire. The wire is galvanized and is a standard size for use with a variety of industrial applications
305 meters of 4/0 stainless steel wire. The wire is 0.8mm diameter. The wire is in a 1.5mm diameter stainless steel tube. The wire is stranded, and is used to make a reel.

##

500 ft of 30 lb. Spectra® nylon monofilament line with a 10' leader and 1/8" braided nylon hook. Spectra® is a trademark of Spectra® LLC.

Item #: 5300UE

2 1/2" x 4" x 500 ft. 3/4" x 1 1/4" x 500 ft. 3/4" x 1 1/2" x 500 ft

500 ft of 0.020" nylon monofilament fishing line. This line is great for all freshwater fishing and is a great choice for tying flies, fishing jigs and other light duty fishing applications.

1,000 ft. of 0.046 in. dia. 304 stainless steel wire rope. For use in general purpose applications.

1,000 ft. of 0.75 mm diameter un-reel color coated steel wire. This wire is suitable for use in any type of putup system.

1,000 ft. of 2.5 lb. per linear foot (2.5 lb./ft.) ReelTuff Box (RBB) with 1/2" diameter, 1/4" wall, 2"

1,000 ft of 1/4" diameter, 30 lb. Green PVC coated wire, ReelTuff Box

Item #: 5300UE 005Z1000, Color: Green, Putup Type: ReelT

1,000 ft of 0.6 mm (24 gauge) solid stainless steel wire rope with a diameter of 0.6 mm (24 gauge) and a length of 1,000 ft. The wire rope is green in color and

1,000 ft of 0.25 in. polypropylene (PP) yarn, 0.25 in. diameter, 10 lb/yd.

1,000 ft of 6.5mm Dia. PVC Coated Wire, 1/4" ID, 1/2" OD, 4 AWG, 20 gauge, 2.5" long, 8.

1,000 ft. of 0.6 mm (0.024") diameter, 0.5 oz. (14 g) glass-fiber yarn, 100% polyester yarn, 100% polyester yarn, 100% polyester yarn

1,000 ft. of 0.10" thick, 2.5 mil polyethylene film. The film is a high density, high clarity, and has a smooth surface finish. The film is UV stable, and is ideal for use
1,000 ft of 0.024" diameter stainless steel wire rope, color white, putup type reel, length
1,000 ft, UPC/EAN: 612825336778, footnotes: C

Item #: 530

1,000 ft of 24 AWG 500V/1000V 6-Conductor PVC Coated Steel Tape

Item #: 5300UE 009U1000, Color: White, Putup Type: UnReel,

1,000 ft of 2" wide polyethylene coated with a polyester coating for a smooth, clean finish. ReelTuff Boxes are made with a tough, durable polyethylene material that resists punctures and tears. The polyester coating makes

1,000 ft. of 0.25mm yellow polyethylene film with a putup of 1.5 meters. The polyethylene is manufactured from high density polyethylene resin and is used for wrapping and bundling. The film is strong

CHATGPT

For comparison, I tested the free version of ChatGPT (accessed directly through the browser). I provided a single row in CSV format and used the same prompt. The output is unedited and raw, with bold words added by the model itself (though this is irrelevant for plain .txt).

```
[',  
'Item #: 5300UE 0101000, Color: Black, Putup Type: Reel, Length:  
1,000 ft, UPC/EAN: 612825159162, Footnote: C']
```

This product, identified by **Item #5300UE 0101000**, is a high-quality black cable supplied in a **reel putup type** for convenient deployment. The **length** of the reel is **1,000 feet**, making it suitable for larger installations and projects. The **UPC/EAN code** is **612825159162**, ensuring easy product identification and tracking. Notably, this item also includes **Footnote C**, which may indicate specific handling or performance characteristics as defined by the manufacturer.

Table 4:

Element	Nom. Conductor DCR	Nom. Capacitance Cond-to-Cond	Max. Current
Conductor(s)	6.5 Ohm/1000ft	21.5 pF/ft (70.5 pF/m)	5
Amps per conductor @ 25°C			

```
[',  
'Element: Conductor(s), Nom. Conductor DCR: 6.5 Ohm/1000ft, Nom. Capacitance  
Cond-to-Cond: 21.5 pF/ft (70.5 pF/m), Max. Current: 5 Amps per conductor @ 25°C']
```

This product features high-quality **conductor elements** engineered to meet specific electrical performance requirements. The **nominal conductor DC resistance (DCR)** is measured at **6.5 Ohms per 1000 feet**, ensuring low resistance over extended lengths for efficient power transmission. The **nominal capacitance between conductors** is **21.5 pF per foot** (or **70.5 pF per meter**), supporting stable electrical properties and minimal signal loss. Each conductor is rated to carry a **maximum current of 5 Amps** at a temperature of **25°C**, making this product ideal for applications requiring reliable current handling and thermal stability.

Table 5:

UL Temperature	Operating	
:-----	:-----	
75°C	-20°C to +75°C	

["UL Temperature: 75°C, Operating: -20°C to +75°C"]

This product is designed to operate reliably across a wide temperature range. It meets the **UL Temperature** standard, rated for up to **75°C**. The **operating temperature** range extends from **-20°C to +75°C**, ensuring performance in both moderately cold and high-temperature environments. These specifications make it suitable for applications where temperature fluctuations are common, providing durability and stability in diverse conditions.

TABLEGPT₂

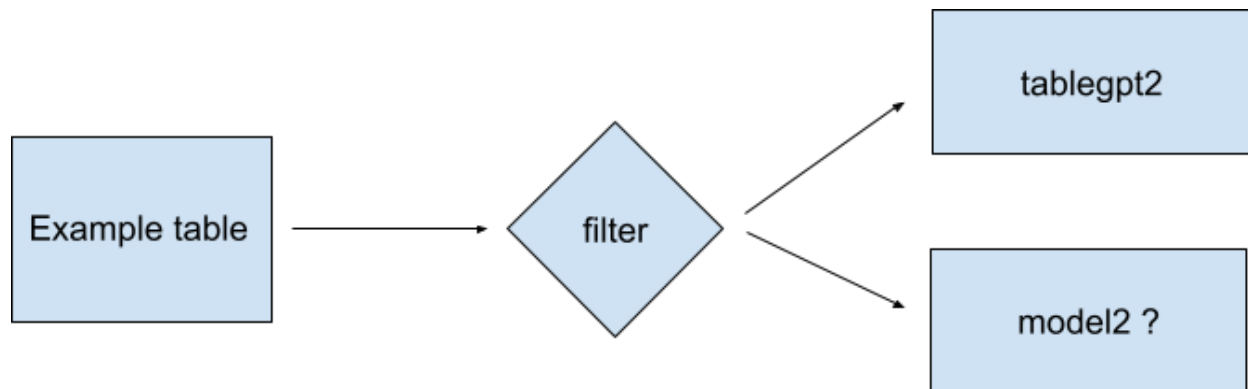
A new model, released in early November, was tested with a focus on table interpretability. The results were generally good, especially compared to other models. However, there are notable limitations:

- The model requires well-structured and properly formatted tables.
- A significant number of "undesired" tables must be filtered out (criteria detailed in the annex). These conditions eliminate about half of the tables we need to process
- Tablegpt2 takes a considerable amount of time to process data (this is acknowledged on the paper also). The example table below took 28 minutes (on my slow gtx 1050 gpu)

Approach Taken

To address this, the following strategy was implemented:

- I already have a filtering function that selects tables eligible for the TableGPT2 model.
- Smaller tables (e.g., those with fewer than five columns and two rows, the largest category filtered out) will be sent to a different model for processing. Based on this report, ChatGPT is the best candidate for this task, as it has demonstrated strong performance with small tables.



Other Options

There are alternative models available, though they are limited, as table-to-text conversion primarily relies on complete, well-structured tables, which do not represent the majority of our dataset.

While exploring additional LLMs such as LLaMA with more parameters or Florence with a larger capacity is an option, I currently lack the necessary **computational resources to proceed**.

Cost Analysis

A detailed cost breakdown for training these tables using the ChatGPT paid API is included in the ANEX

Table 6: Tested using tablegpt2, table included in csv format

[',
'Item #: 5300FE 0101000, Color: Black, Putup Type: Reel, Length: 1,000 ft, UPC/EAN: 612825158981',
'Item #: 5300FE 010U1000, Color: Black, Putup Type: UnReel, Length: 1,000 ft, UPC/EAN: 612825423362',
'Item #: 5300FE 010Z1000, Color: Black, Putup Type: ReelTuff Box, Length: 1,000 ft, UPC/EAN: 612825159025',
'Item #: 5300FE 0061000, Color: Blue, Light, Putup Type: Reel, Length: 1,000 ft, UPC/EAN: 612825350323',
'Item #: 5300FE.00152, Color: Chrome, Putup Type: Reel, Length: 152 m, UPC/EAN: 8719605008145',
'Item #: 5300FE.00305, Color: Chrome, Putup Type: Reel, Length: 305 m, UPC/EAN: 8719605008152',
'Item #: 5300FE 008500, Color: Gray, Putup Type: Reel, Length: 500 ft, UPC/EAN: 612825336754',
'Item #: 5300FE 008U500, Color: Gray, Putup Type: UnReel, Length: 500 ft, UPC/EAN: 612825423348',
'Item #: 5300FE 0081000, Color: Gray, Putup Type: Reel, Length: 1,000 ft, UPC/EAN: 612825158974, Footnote: C',
'Item #: 5300FE 008U1000, Color: Gray, Putup Type: UnReel, Length: 1,000 ft, UPC/EAN: 612825158950',
'Item #: 5300FE 008Z1000, Color: Gray, Putup Type: ReelTuff Box, Length: 1,000 ft, UPC/EAN: 612825350330',
'Item #: 5300FE 0031000, Color: Orange, Putup Type: Reel, Length: 1,000 ft, UPC/EAN: 612825350309',
'Item #: 5300FE 0021000, Color: Red, Putup Type: Reel, Length: 1,000 ft, UPC/EAN: 612825423331',
'Item #: 5300FE 002Z1000, Color: Red, Putup Type: ReelTuff Box, Length: 1,000 ft, UPC/EAN: 612825350293',
'Item #: 5300FE 0071000, Color: Violet, Putup Type: Reel, Length: 1,000 ft, UPC/EAN: 612825158967',
'Item #: 5300FE 0091000, Color: White, Putup Type: Reel, Length: 1,000 ft, UPC/EAN: 612825423355',
'Item #: 5300FE 009Z1000, Color: White, Putup Type: ReelTuff Box, Length: 1,000 ft, UPC/EAN: 612825158998',

'Item #: 5300FE 0041000, Color: Yellow, Putup Type: Reel, Length: 1,000 ft, UPC/EAN: 612825350316']

General Overview:

- **Colors Available:** Black, Light Blue, Chrome, Gray, Orange, Red, Violet, White, and Yellow.
- **Putup Types:** Reel, UnReel, and ReelTuff Box.
- **Lengths:** Primarily 1,000 feet or 500 feet, with some items measured in meters (152 m and 305 m).
- **UPC/EAN Codes:** Each item has a unique UPC/EAN for inventory and identification.

Item Breakdown:

1. **Black Items**
 - **Item #5300FE 0101000:** Black, 1,000 ft, Reel, UPC/EAN 612825158981
 - **Item #5300FE 010U1000:** Black, 1,000 ft, UnReel, UPC/EAN 612825423362
 - **Item #5300FE 010Z1000:** Black, 1,000 ft, ReelTuff Box, UPC/EAN 612825159025
2. **Light Blue Item**
 - **Item #5300FE 0061000:** Light Blue, 1,000 ft, Reel, UPC/EAN 612825350323
3. **Chrome Items**
 - **Item #5300FE.00152:** Chrome, 152 m, Reel, UPC/EAN 8719605008145
 - **Item #5300FE.00305:** Chrome, 305 m, Reel, UPC/EAN 8719605008152
4. **Gray Items**
 - **Item #5300FE 008500:** Gray, 500 ft, Reel, UPC/EAN 612825336754
 - **Item #5300FE 008U500:** Gray, 500 ft, UnReel, UPC/EAN 612825423348
 - **Item #5300FE 0081000:** Gray, 1,000 ft, Reel, UPC/EAN 612825158974, Footnote: C
 - **Item #5300FE 008U1000:** Gray, 1,000 ft, UnReel, UPC/EAN 612825158950
 - **Item #5300FE 008Z1000:** Gray, 1,000 ft, ReelTuff Box, UPC/EAN 612825350330
5. **Orange Item**
 - **Item #5300FE 0031000:** Orange, 1,000 ft, Reel, UPC/EAN 612825350309
6. **Red Items**
 - **Item #5300FE 0021000:** Red, 1,000 ft, Reel, UPC/EAN 612825423331
 - **Item #5300FE 002Z1000:** Red, 1,000 ft, ReelTuff Box, UPC/EAN 612825350293
7. **Violet Item**
 - **Item #5300FE 0071000:** Violet, 1,000 ft, Reel, UPC/EAN 612825158967
8. **White Items**
 - **Item #5300FE 0091000:** White, 1,000 ft, Reel, UPC/EAN 612825423355
 - **Item #5300FE 009Z1000:** White, 1,000 ft, ReelTuff Box, UPC/EAN 612825158998
9. **Yellow Item**
 - **Item #5300FE 0041000:** Yellow, 1,000 ft, Reel, UPC/EAN 612825350316

Notes

- **Item Number Patterns:** The item numbers typically follow a similar pattern with variations for color, length, and packaging style.
- **Footnote "C":** Only the Gray item #5300FE 0081000 has a footnote labeled "C," which may indicate a special characteristic or note associated with that item.

This detailed description categorizes the items based on their unique identifiers and attributes for easy reference

Conclusion

To proceed, I need clear direction on the following:

1. Model Selection and Workflow Validation

- Which model should be used, and does the current workflow seem appropriate?
- If testing additional LLMs is recommended, I would need suggestions for potential models (based on experience or relevance) and, critically, access to a server due to the computational limitations of my laptop.

2. ChatGPT API Usage

- If the ChatGPT API is the preferred approach, I would need permission to utilize it and guidance on how to proceed.

Remaining Tasks for Table-to-Text Workflow

To finalize the table-to-text workflow (from reading tables in PDFs to processing, filtering, calling LLMs, and writing results to files), the final step is incorporating the second LLM model into the existing code.

Known Limitations

- **Data Duplication:** We currently duplicate information by processing the document as "text" and including everything, then separately processing tables and adding the generated text. Bookmarking locations was tested but did not work. However, this is unlikely to cause significant issues, as the overall data volume is manageable.
- **Schedule Delays:** We are behind schedule, as the full dataset still needs to be converted into plain text. While text processing is not an issue, completing table processing requires the steps outlined here.

Image Processing Workflow

Although image processing falls outside the scope of this report, key tasks to complete include:

1. **Training a Classifier:** A preliminary version achieves ~95% accuracy, but I need to label 50,000 images to finalize the dataset. The preliminary version was trained using a free service that only categorizes up to 10,000 images/month. The paid plan is way too expensive. Need to retrain the model on our own hardware
2. **Incorporating the Classifier:** The classifier must be integrated into the workflow once finalized.
3. **Testing a Larger Model:** The current LLaMA model for image explanations is decent but constrained by my limited hardware. Testing a larger model would be beneficial.

Final Requirement

To complete these tasks, especially testing additional LLMs or larger models, **I require access to robust computational resources.**

Next Steps

In the meantime I will start looking into VSM models (which I believe to be the next step of the process). If there is any guidance you can give on this topic it would be greatly appreciated, i already have a literature group of documents to review so i'll start reading those.

TABLE PREPROCESSING

	Applicable limit switches	Lamp connection	Lamp rating	Part No.
0	Applicable limit switches	Lamp connection	Lamp rating	Part No.
1	Exposed terminal types	Normally open connection	6V DC	None
2	None	None	12V DC	None
3	None	None	24V to 48V DC	None
4	None	Normally closed connection	6V DC	None
5	None	None	12V DC	None
6	None	None	24V to 48V DC	None

-----0

	Lamp rating
0	6V DC
1	12V DC
2	24V to 48V DC
3	6V DC
4	12V DC
5	24V to 48V DC

ANEX

Selection conditions for table processing in tablegpt2.

copied straight from the tablegpt2 paper <https://arxiv.org/abs/2411.02059>

Table Selection and Column Manipulation

To ensure high-quality and consistent data for the BI application scenarios, we apply a systematic set

of rules for table filtering and selection. Below are the key filtering rules we follow:

- remove horizontal tables: Any tables structured horizontally are transposed or removed to maintain consistency in data orientation.
- remove duplicate columns or columns ignoring case sensitivity: Columns with identical names or names that differ only by case (e.g., “Column” and “column”) are merged or discarded.
- remove columns containing only underscores: Columns that only contain underscores or similar meaningless data are removed.
- remove columns where field content exceeds 100 characters: Columns containing overly long text fields, which may be irrelevant or quite uncommon realistically, are filtered out.
- remove rows or columns with more than 30% NaN Values: Any tables where a significant portion of the data is missing are filtered out to ensure data completeness.

- remove tables with fewer than 5 rows or 2 columns: Tables with insufficient data (less than 5 rows or 2 columns) are considered too sparse and are discarded.
- remove columns where the first value directly matches the column name: When a column's first entry repeats its own column name, it's a sign of an error or redundant information, so the column is removed.
- more to be omitted in this report...

These rules ensure that tables used in our model training are relevant, clean, and structured, optimizing the quality of the data pipeline for better performance.

After reviewing the ChatGPT API pricing, I've estimated the cost for describing tables extracted from our documents. Here's the breakdown of my calculations:

1. **Batch Processing Discount:** By opting for 24-hour batch processing (50% discount), we can get 1 million tokens for \$1.25. Since 1,000 tokens cover approximately 750 words, we can process 750,000 words per \$1.25.
2. **Token Usage per Table:**
 - **Single-row tables:** Based on our samples, about 80% of tables contain only one row, requiring approximately 70 words to describe each.
 - **Multi-row tables:** The remaining 20% of tables average 10 rows. A prompt that generates a detailed description for each row produces 20–50 words per row, or 200–500 words per table.
3. Based on this, the average word count needed per table is 96–156 words.
4. **Words per Document:** With around 10 tables per document, the average document will require 960–1,560 words for table descriptions.
5. **Total Tokens Needed:**
 - For 3,000 documents, we would need approximately 2.88–4.68 million words (best and worst-case estimates). This translates to 3.84–6.24 million tokens, costing around 5\$–7.50\$.
 - Adding a 50% buffer for potential error, the final cost estimate is 7.50\$–12.00\$.

Please let me know if these calculations align with your expectations or if any adjustments are needed.

Please note that these calculations are estimates based on the entire dataset. The actual word and token count will be lower after filtering data for TableGPT processing.