Comet

Only Prop analysis

- Wrong entity extracted as given Property: 29.5%
- Composition Not shared/ not-relevant to the paper: 9.09%
- Present in Table itself: 20.45%
- Present in Tables description: 11.36%
- Present in text: 31.81%

Where can composition be present in the table?

- In the table itself.
  - With ID, Without any ID defined.
- Title of the paper.
- In table’s description.
- In the abstract.
- In the introduction
- In the methodology
- In results
- In Text w ID
- Equational/Direct

Only Prop Tables

Redo:

- – Present in the Table itself: 46.67 %
- – ID Present in tables: 60%
- – Present in table’s Description: 53%
- **– Title of the paper: 73.33%**
- – Abstract: 73.33% , Introduction: 66.67% , Methodology: 80%, Results: 66.67%
- – ID in text: 60%
- – Equational Composition: 40%

Only Prop Tables

– Present in the Table itself: 26.53 %

– Title of the paper: 18.36%

– Abstract: 26.53%

– Introduction: 28.57%

– Methodology: 30.61%

– Results, Discussions, Conclusion: 28.57%

Types of papers without proper compositions

Processing conditions

Composites and not chemical compounds.

To connect Properties in Tables with compositions in text

- Use/Get baseline performance using LLMs
- Modify Dhruvil’s architecture to handle material IDs too.
- A new pipeline inspired from Dhruvil’s architecture.

Baseline Performance of VLLMs

Provide the text from the research paper with images of extracted table from Kausik’s architecture with the composition field’s as empty/fill in the blank.

Modified Dhruvil’s architecture

Finetune the DC and EC model’s to also extract material IDs, which we can directly correlate with the table’s.

A new pipeline, using Small Vision LMs

Finetuned on take in sentences containing the composition of materials along with an image of the table generated by Kausik’s architecture to fill in the gaps in terms of missing compositions.

Suggested model - phi3.5-vision Mini

128,000 token context window

3.8 B parameters.

Only Prop Tables

Redo:

- Present in the Table itself: 46.67 %
- ID Present in tables: 60%
- Present in table’s Description: 53%
- **Title of the paper: 73.33%**
- Abstract: 73.33% , Introduction: 66.67% , Methodology: 80%, Results: 66.67%
- ID in text: 60%
- Equational Composition: 40%

Only Prop Tables

– Present in the Table itself: 26.53 %

– Title of the paper: 18.36%

– Abstract: 26.53%

– Introduction: 28.57%

– Methodology: 30.61%

– Results, Discussions, Conclusion: 28.57%

But this is an incomplete picture!

Strategy 1 - Prompting a LLM

- Single-Shot prompting w/ complete context:
  - System Prompt
  - Instructions
  - Example Input: Research Paper + Incomplete Table + Example Output
  - Target Research Paper + Incomplete Tables

System Prompt

“You are a material scientist assisting researchers in extracting and structuring scientific data from research papers. I have extracted some properties from the tables in the research paper, but unable to extract the corresponding compositions. Your task is to read the provided research paper and use it to extract the missing compositions and complete the table. “

Single-Shot prompt w/ Complete context

Instructions:

1. Input Format: You will receive:

   - A research paper text.

   - An incomplete table with placeholders like `<blank_1>`, `<blank_2>`, etc. for the missing compositions.

2. Task: Identify and extract the relevant compositions from the research paper to fill in the placeholders in the table. Ensure that:

   - The extracted information is precise and relevant to the placeholders in the table.

   - The table is filled based on the context provided in the research paper and the incomplete table.

3. Output Format: Return the list of blanks and their corresponding compositions.

Context = Research Paper +

Incomplete Table:

```
    Article PII Table No. ID          Proxy_ID Composition             Property      Journal_Name
S0167577X12003023     [1]    S0167577X12003023_0_R_2   <blank_1> ('Poisson ratio', 0.2438, None) Materials_Letters
S0167577X12003023     [1]    S0167577X12003023_0_R_3   <blank_2> ('Poisson ratio', 0.2604, None) Materials_Letters
```

Example Output:

<blank_1> = CA/PVP-Amoxi/CA at pH 7.2

<blank_2> = CA/PVP-Amoxi/CA at pH 3.0

Models Prompted-

- Gemini-1.5-flash-latest
  - Input token limit - 1,048,576
  - Output token limit - 8,192
  - Free:
  - 15 RPM
  - 1 million TPM
  - 1,500 RPD

Gemini-1.5-Flash-latest as of 31st December 2024

Maximum token count: 49123

Minimum token count: 8766

Maximum processing time: 25.649644136428833

Minimum processing time: 2.675307035446167

Total token count: 2071531

Total processing time: 533.0484998226166 s

Results

Correct - 57.60%

Total - 92 unique pii’s

Predictions -

- Prompt-Quality Improvements

Bottlenecks with current prompting strategy

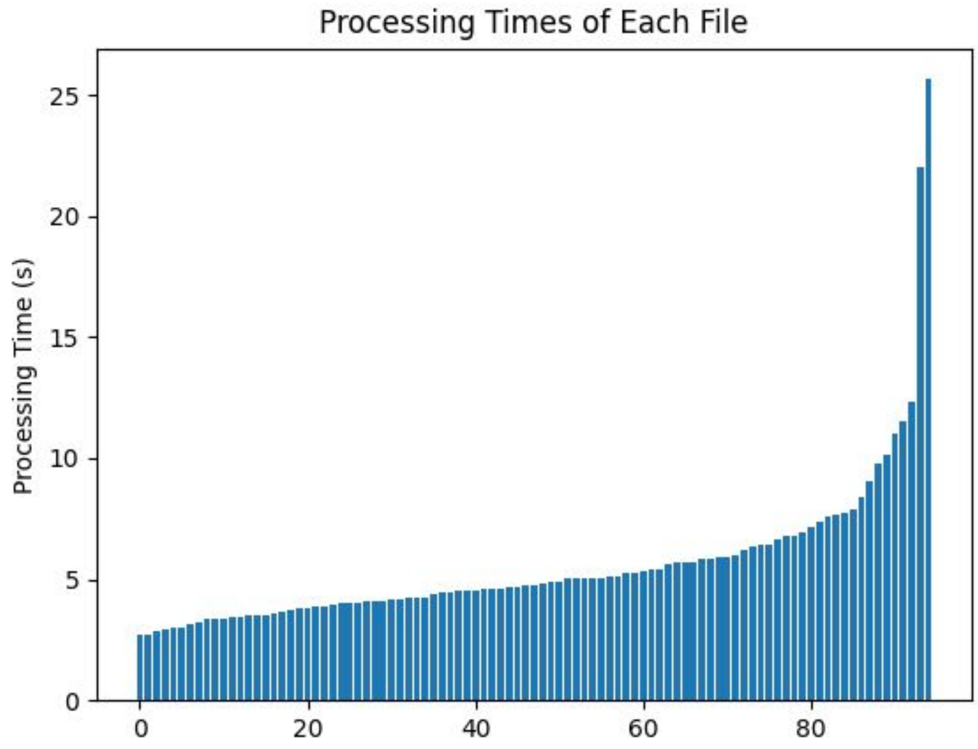
Unscalable

Computationally Expensive

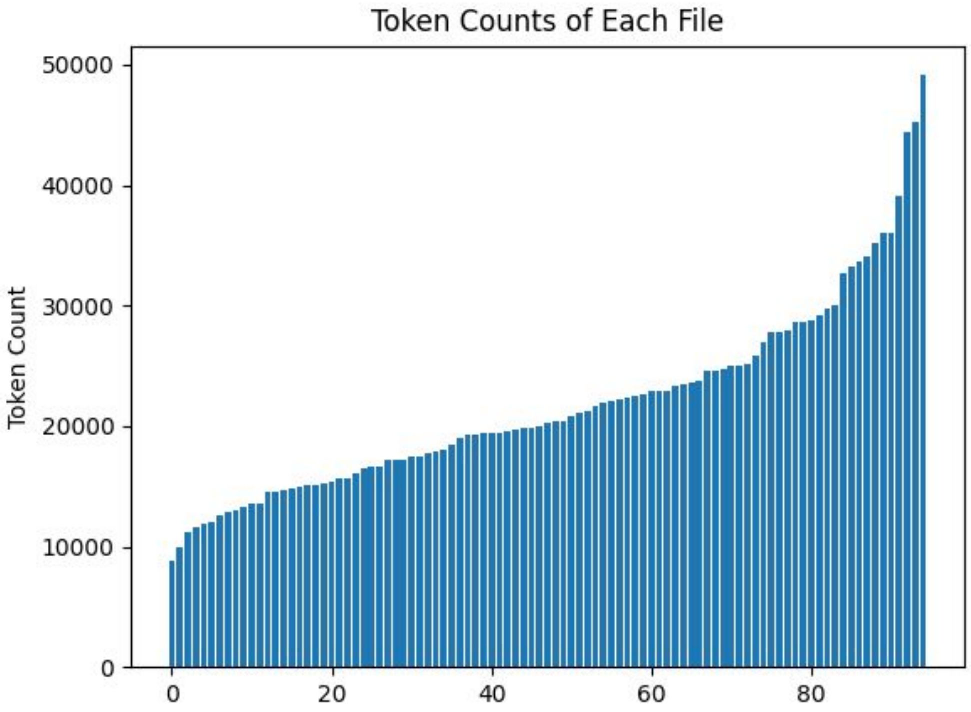
API Costs too steep

Runtime High

Runtime Analysis



Token Count Analysis



Cost Analysis

Gpt-4o -

$2.5/1M Input tokens w/o Caching

For 95 papers: ~5.178 USD

For ~50000 papers/tables: ~2,725.6986842105 USD

$1.25/1M Input tokens w Caching

For 95 papers: ~1.920495 USD

For ~50000 papers/tables: ~1,010.786842 USD

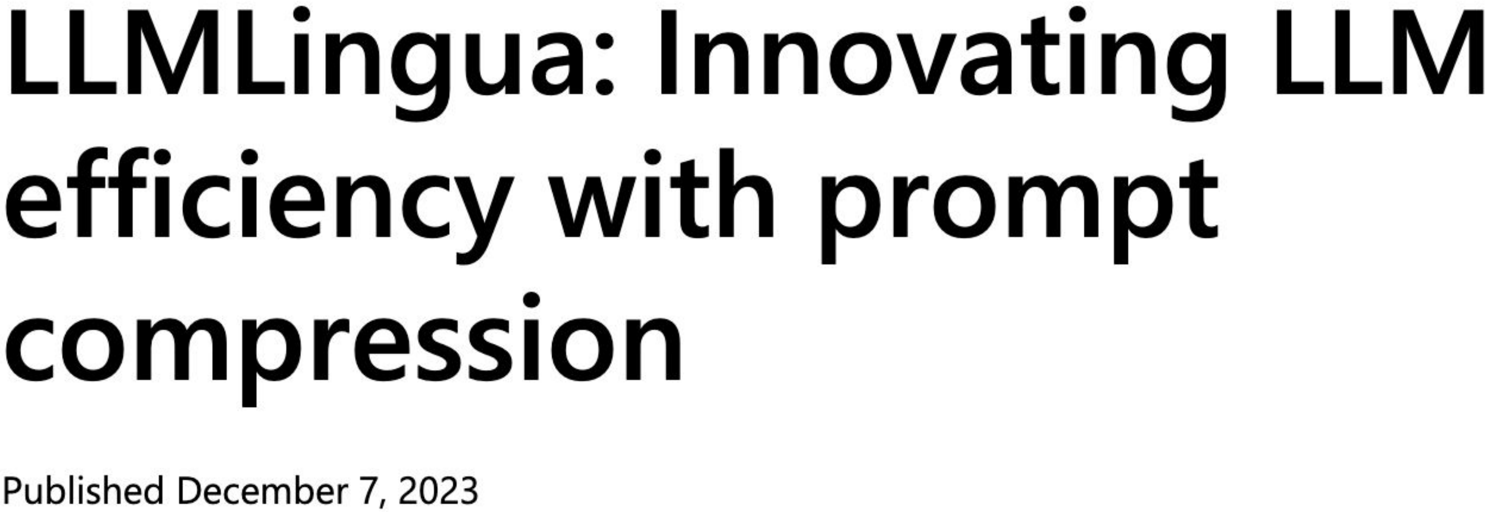
Future Ideas For LLMs

- Prompt Compression
- Improved Prompting

Prompt Compression

Use LLMLingua?

Results - Delayed a bit! (Next week for sure!)



LLMLingua: Innovating LLM efficiency with prompt compression

Published December 7, 2023

Original Prompt(9-steps Chain-of-Thought):

Question: Sam bought a dozen boxes, each with 30 highlighter pens inside, for $10 each box. He rearranged five of these boxes into packages of six highlighters each and sold them for $3 per package. He sold the rest of the highlighters separately at the rate of three pens for $2. How much profit did he make in total, in dollars?

Let's think step by step
Sam bought 12 boxes x $10 = $120 worth of highlighters.
He bought 12 * 30 = 360 highlighters in total.
Sam then took 5 boxes × 6 highlighters/box = 30 highlighters.
He sold these boxes for 5 * $3 = $15
After selling these 5 boxes there were 360 - 30 = 330 highlighters remaining.
These form 330 / 3 = 110 groups of three pens.
He sold each of these groups for $2 each, so made 110 * 2 = $220 from them.
In total, then, he earned $220 + $15 = $235.
Since his original cost was $120, he earned $235 - $120 = $115 in profit.
The answer is 115

Compressed Prompt:

: Sam bought a dozen boxes each 30 highl pens inside, $10 each. He reanged five of boxes into of six each $3 per. He sold the thelters separately at the of three $2. much make total,
Lets think step
bought boxes x0 oflters
He 2 3ters in
Sam then boxes 6lters/box 0ters
He sold these boxes 5
Afterelling these boxes there 36030lters
ese00 of three
sold groups2 each so made *2 $20 from
In total, he015
Since his he $ - $120 = $115 in profit.
The answer is 115

Problem Statement ???

Currently where are LLMs mostly failing apart from obvious?

- In Papers which are not conventional Composition-Property for a simple-material but rather complex situations/structures.
- Composite Material
- Doped Material
- Opto-thermal thin films
- Layers
- Opto-thermal thin films
- etc

Only Prop Tables

– Present in the Table itself: 26.53 %

– Title of the paper: 18.36%

– Abstract: 26.53%

– Introduction: 28.57%

– Methodology: 30.61%

– Results, Discussions, Conclusion: 28.57%

But this is an incomplete picture!

Only Prop Tables

Classical Chemical composition not defined =>

61.22%

Composition Well Defined => accuracy = 73.68% (will rise to ~90+% with tailored improved prompting strategy)

For scalability, easily a small model architecture can be developed!

Composition Undefined/ Not Well Defined => accuracy = 42.30%

We will have to define this properly! What do we want to extract.

Failure Modes?

What I have found yet in the research papers?

- Variation in computational methods

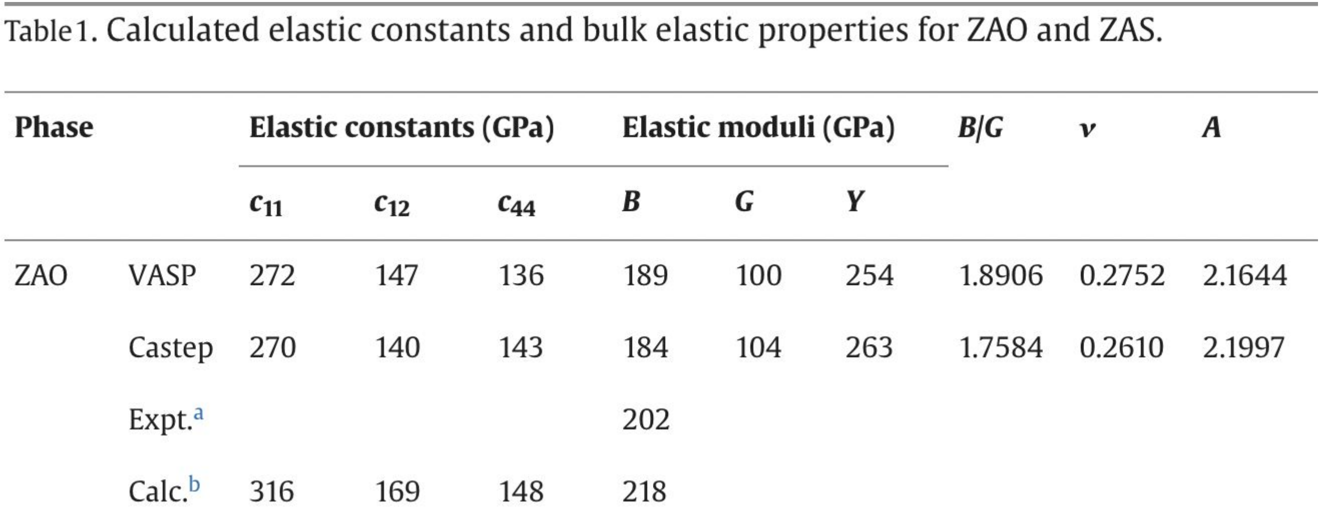


Table 1. Calculated elastic constants and bulk elastic properties for ZAO and ZAS.

Phase		Elastic constants (GPa)			Elastic moduli (GPa)			*B*/*G*	*ν*	*A*
		c_{11}	c_{12}	c_{44}	*B*	*G*	*Y*			
ZAO	VASP	272	147	136	189	100	254	1.8906	0.2752	2.1644
	Castep	270	140	143	184	104	263	1.7584	0.2610	2.1997
	Expt.[a]				202					
	Calc.[b]	316	169	148	218					

S1293255814002568

Failure Modes?

- Varying Synthesis process conditions

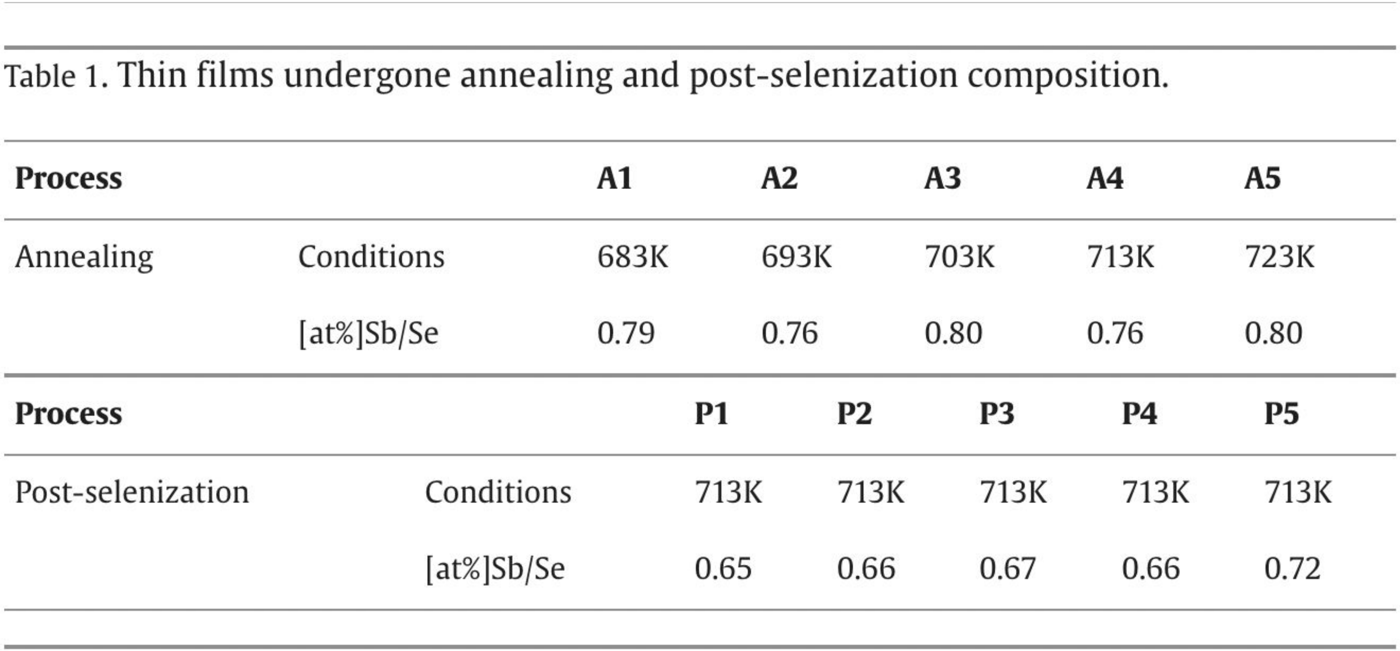


Table 1. Thin films undergone annealing and post-selenization composition.

Process		**A1**	**A2**	**A3**	**A4**	**A5**
Annealing	Conditions	683K	693K	703K	713K	723K
	[at%]Sb/Se	0.79	0.76	0.80	0.76	0.80
Process		**P1**	**P2**	**P3**	**P4**	**P5**
Post-selenization	Conditions	713K	713K	713K	713K	713K
	[at%]Sb/Se	0.65	0.66	0.67	0.66	0.72

S0272884223026500

Failure Modes?

- Varying the raw material for synthesis
    - In the processing conditions

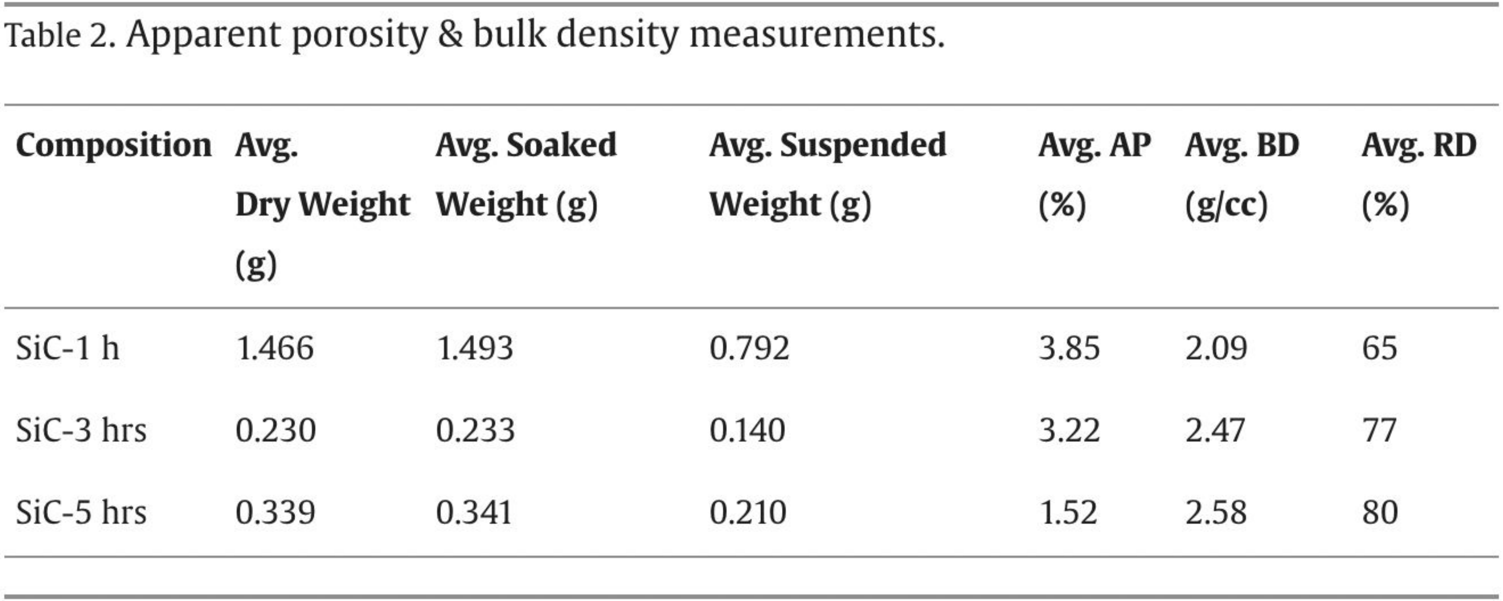


Table 2. Apparent porosity & bulk density measurements.

Composition	Avg. Dry Weight (g)	Avg. Soaked Weight (g)	Avg. Suspended Weight (g)	Avg. AP (%)	Avg. BD (g/cc)	Avg. RD (%)
SiC-1 h	1.466	1.493	0.792	3.85	2.09	65
SiC-3 hrs	0.230	0.233	0.140	3.22	2.47	77
SiC-5 hrs	0.339	0.341	0.210	1.52	2.58	80

S0272884223003978

Failure Modes?

- Varying the raw material for synthesis
    - In the raw materials proportions

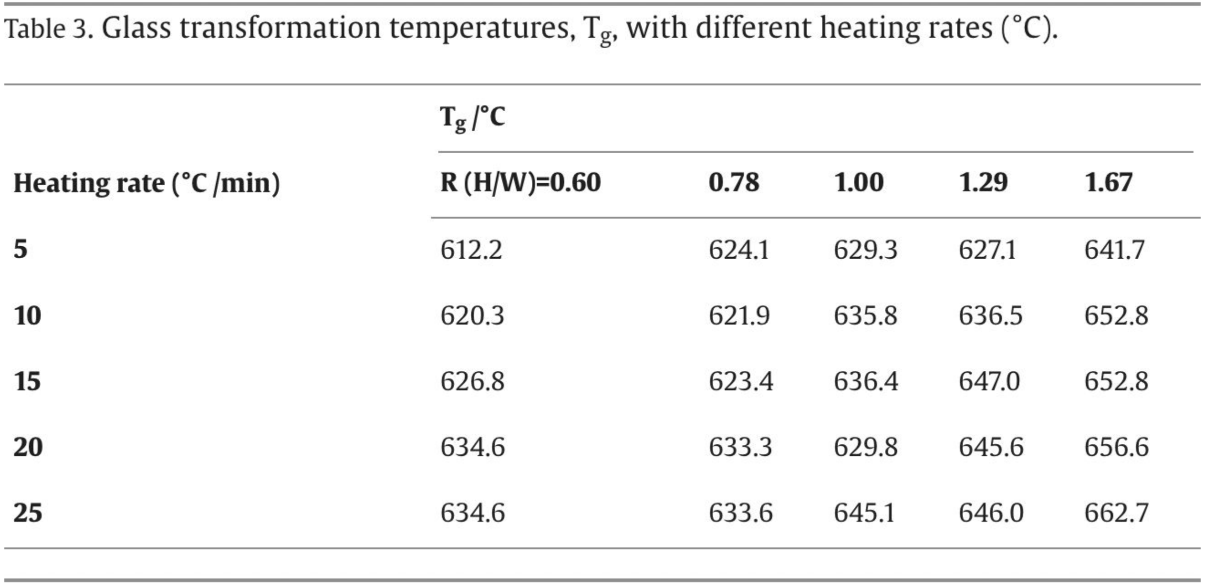


Table 3. Glass transformation temperatures, T_g, with different heating rates (°C).

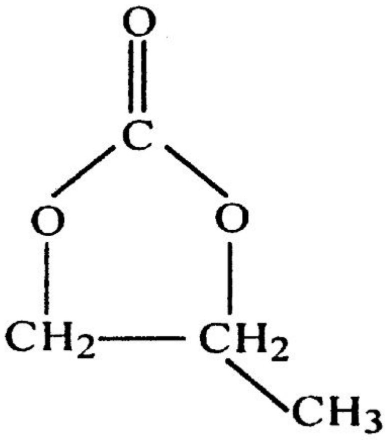
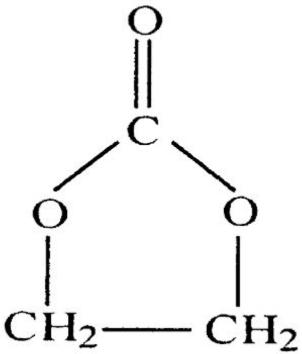
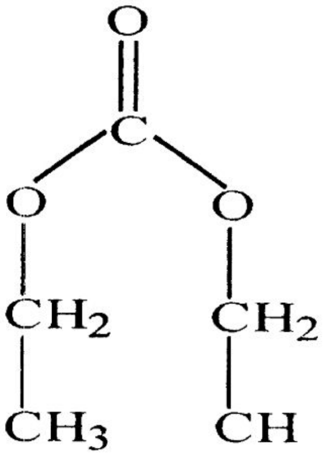
	T_g /°C				
Heating rate (°C /min)	**R (H/W)=0.60**	**0.78**	**1.00**	**1.29**	**1.67**
5	612.2	624.1	629.3	627.1	641.7
10	620.3	621.9	635.8	636.5	652.8
15	626.8	623.4	636.4	647.0	652.8
20	634.6	633.3	629.8	645.6	656.6
25	634.6	633.6	645.1	646.0	662.7

high-carbon ferrochromium slag (HCFS) : Waste Glass (W)

S0272884216316327

Failure Modes?

- Organic Structures

Characteristic	PC[a]	EC[a]	DEC[a]
Structural formula			

S0167273800007554

Influence of deposition temperature on microstructure and electrical properties of modified (Ba, Sr)TiO_3 ferroelectric thin films

S0272884215005635

The impact toughness characteristics of steel wire-reinforced polymer composites

S0167577X99000282

Where are Matskraft and LLMs both failing?

- Nanoparticles synthesis
- Effect of Synthesis conditions on properties
- Effect of Initial composition/Raw materials on Synthesis
- Doping
- Thin Film growth
- Effect of Experimental conditions on properties
- Composition Defined in terms of common names and not chemical formulas
- Final composition not defined in most of these cases.

Solution?

Replace composition with “description of the material”?

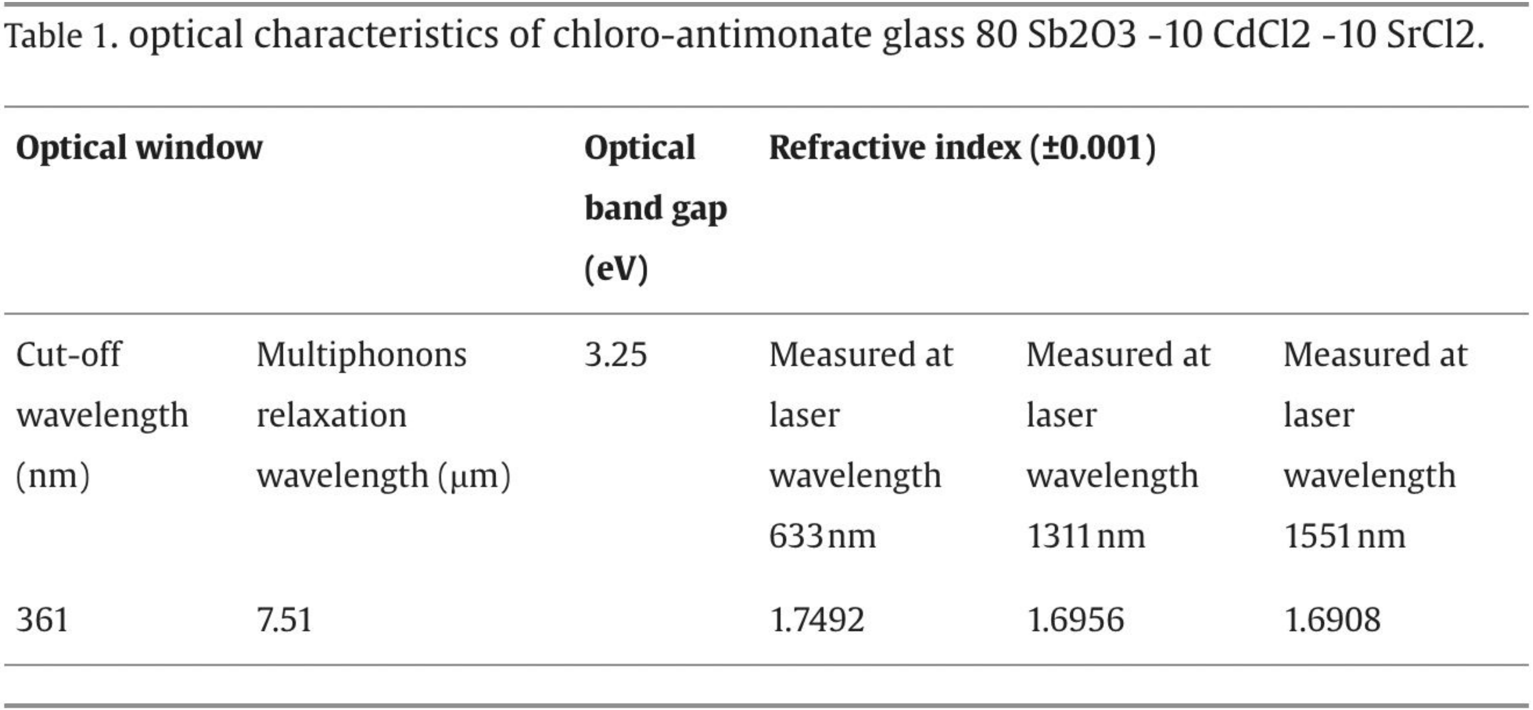
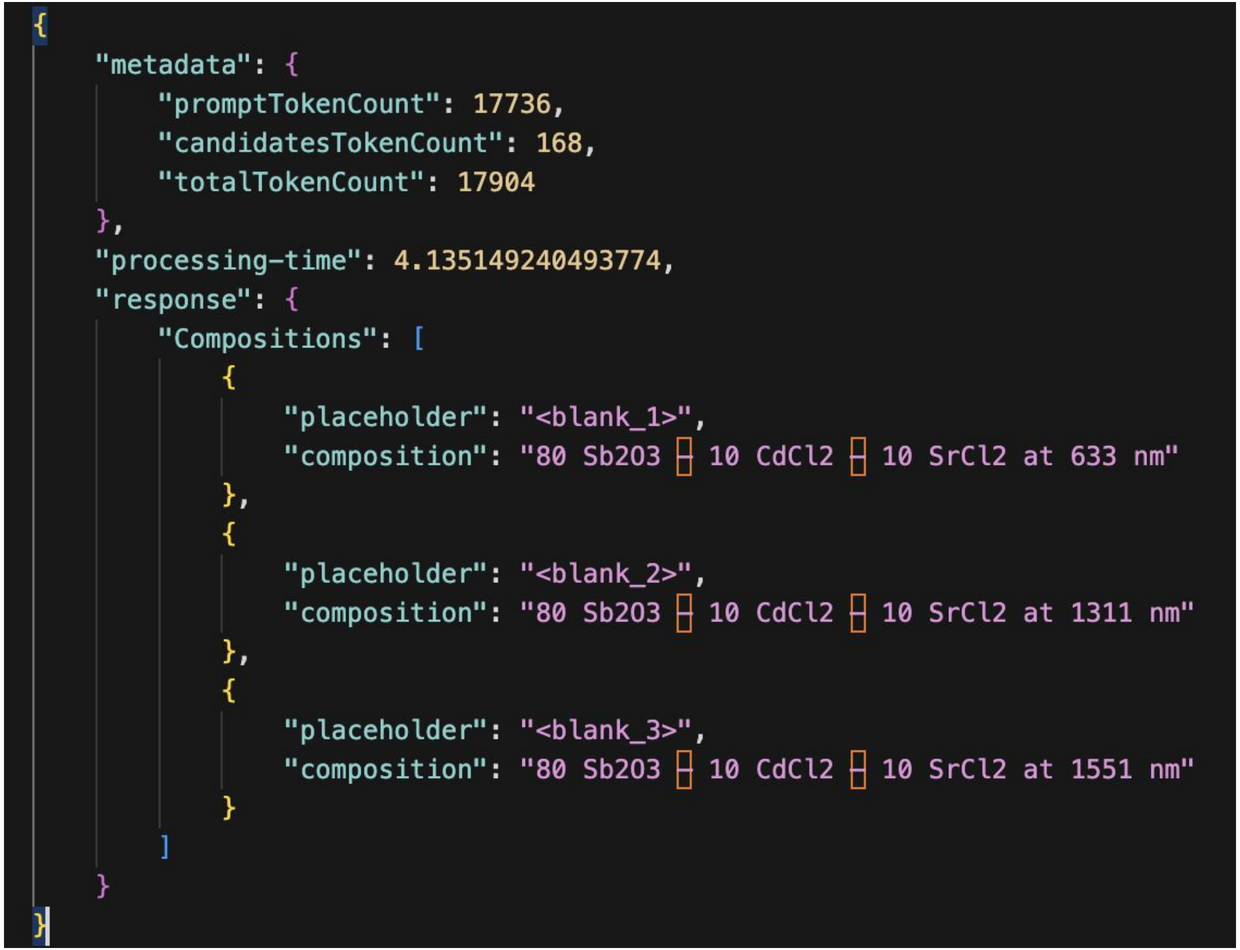


Table 1. optical characteristics of chloro-antimonate glass 80 Sb2O3 -10 CdCl2 -10 SrCl2.

Optical window		**Optical band gap (eV)**	**Refractive index (±0.001)**		
Cut-off wavelength (nm)	Multiphonons relaxation wavelength (μm)	3.25	Measured at laser wavelength 633 nm	Measured at laser wavelength 1311 nm	Measured at laser wavelength 1551 nm
361	7.51		1.7492	1.6956	1.6908



```
{
    "metadata": {
        "promptTokenCount": 17736,
        "candidatesTokenCount": 168,
        "totalTokenCount": 17904
    },
    "processing-time": 4.135149240493774,
    "response": {
        "Compositions": [
            {
                "placeholder": "<blank_1>",
                "composition": "80 Sb2O3 □ 10 CdCl2 □ 10 SrCl2 at 633 nm"
            },
            {
                "placeholder": "<blank_2>",
                "composition": "80 Sb2O3 □ 10 CdCl2 □ 10 SrCl2 at 1311 nm"
            },
            {
                "placeholder": "<blank_3>",
                "composition": "80 Sb2O3 □ 10 CdCl2 □ 10 SrCl2 at 1551 nm"
            }
        ]
    }
}
```

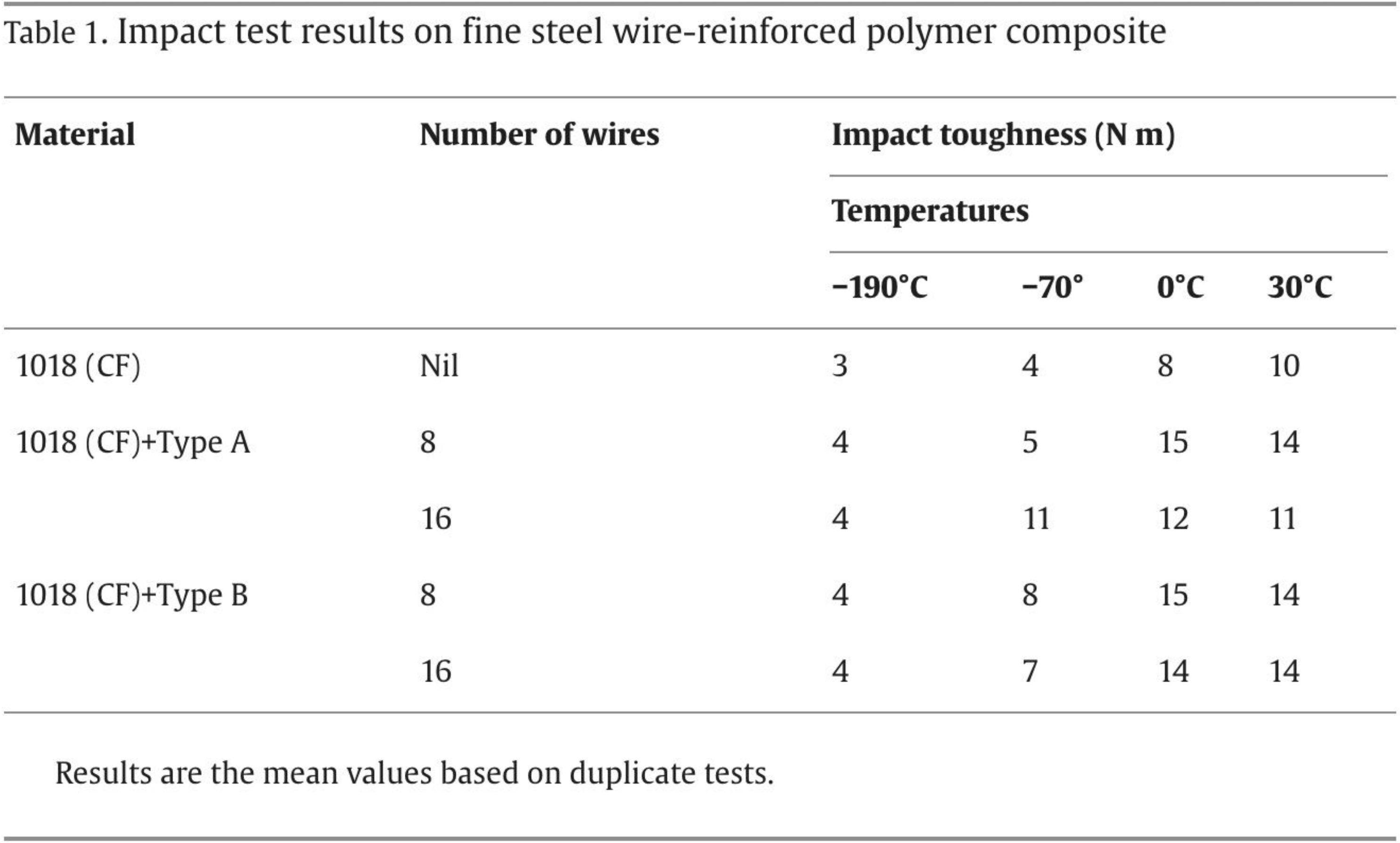
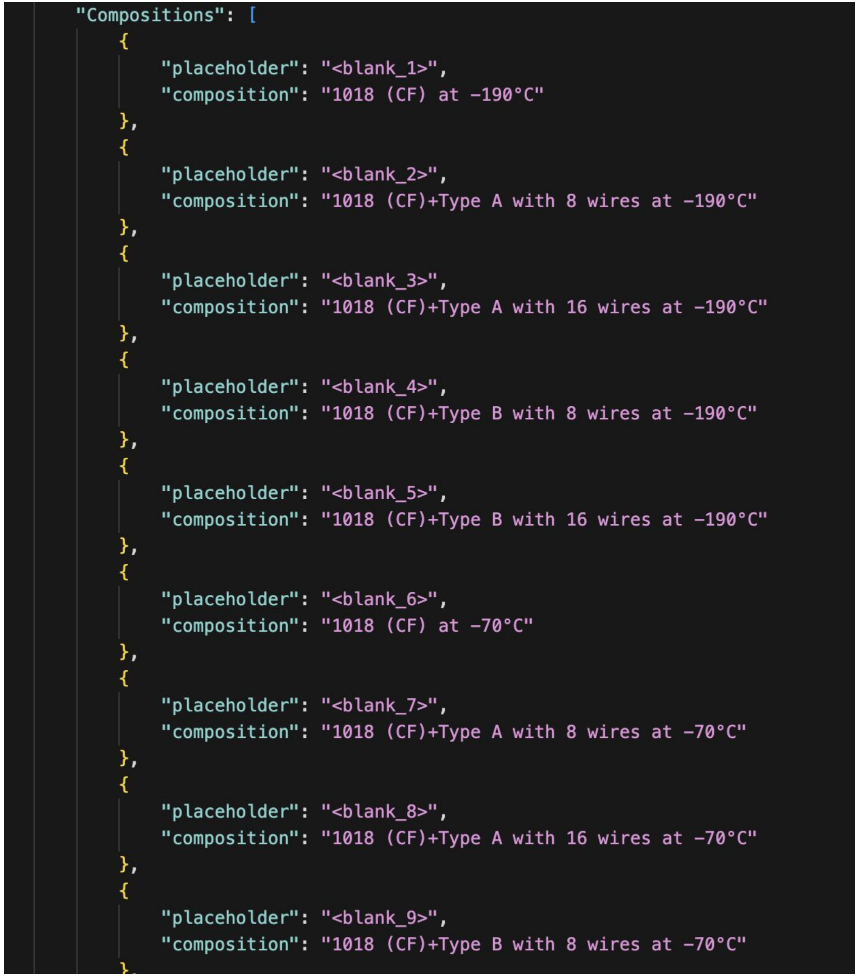


Table 1. Impact test results on fine steel wire-reinforced polymer composite

Material	Number of wires	Impact toughness (N m)			
		Temperatures			
		−190°C	−70°	0°C	30°C
1018 (CF)	Nil	3	4	8	10
1018 (CF)+Type A	8	4	5	15	14
	16	4	11	12	11
1018 (CF)+Type B	8	4	8	15	14
	16	4	7	14	14

Results are the mean values based on duplicate tests.



```
        "Compositions": [
            {
                "placeholder": "<blank_1>",
                "composition": "1018 (CF) at −190°C"
            },
            {
                "placeholder": "<blank_2>",
                "composition": "1018 (CF)+Type A with 8 wires at −190°C"
            },
            {
                "placeholder": "<blank_3>",
                "composition": "1018 (CF)+Type A with 16 wires at −190°C"
            },
            {
                "placeholder": "<blank_4>",
                "composition": "1018 (CF)+Type B with 8 wires at −190°C"
            },
            {
                "placeholder": "<blank_5>",
                "composition": "1018 (CF)+Type B with 16 wires at −190°C"
            },
            {
                "placeholder": "<blank_6>",
                "composition": "1018 (CF) at −70°C"
            },
            {
                "placeholder": "<blank_7>",
                "composition": "1018 (CF)+Type A with 8 wires at −70°C"
            },
            {
                "placeholder": "<blank_8>",
                "composition": "1018 (CF)+Type A with 16 wires at −70°C"
            },
            {
                "placeholder": "<blank_9>",
                "composition": "1018 (CF)+Type B with 8 wires at −70°C"
            },
```

Key Failure Modes

1. Incomplete Data Extraction:

fragmented or implicit compositions (e.g., ratios without formulas, qualitative descriptions).

2. Format Limitations:

Figures are not parsed effectively

3. Contextual Blind Spots:

Compositions tied to processing (e.g., "sintered at different temperatures") or composites (e.g., "fibers embedded in a matrix") are missed.

4. Overfitting to Explicit Formulas:

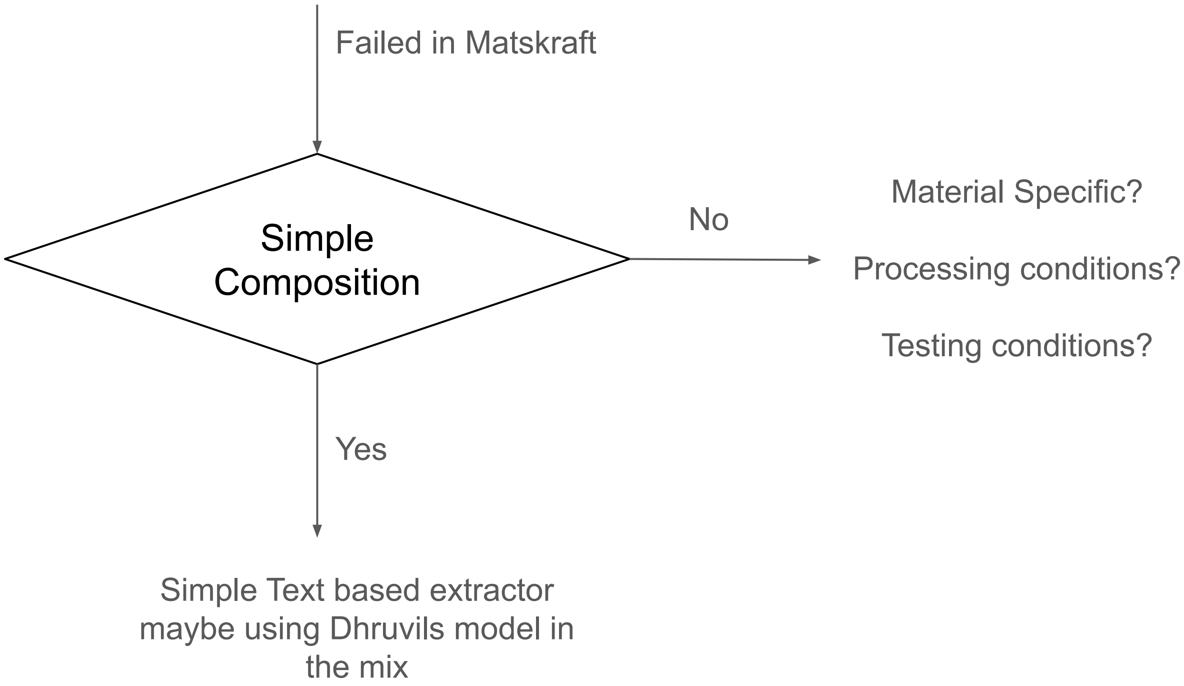
standard chemical formulas (e.g., Fe_2O_3) but fails for non-standard representations (e.g., "olive oil suspension")

Bigger Question: HOW?

- Material Specific Prompting/FT-ing!!!

- Classify Materials first -
  - Metals/Alloys
  - Ceramics
  - Polymers
  - Composites

- Classify Properties - (based on how its defined in the paper)
  - Environment dependent
  - Environment independent



Failed in Matskraft

Simple Composition

No

Material Specific?

Processing conditions?

Testing conditions?

Yes

Simple Text based extractor maybe using Dhruvils model in the mix

4th March

Quick Updates

By the end of this week, build:

- 3 Version ZERO’s for material information extraction using LLMs
- V 0.0.1 - Composition Extraction and Matching
- V 0.0.2 - Property Extraction and Matching
- V 0.0.3 - Both common extraction - ?

Version zero.0’s ?

- Use only LLM (Gemini-2.0/GPT-4o) Cleaned/Generated Training Data
- Instruction Fine-tuning
- Model?

Why the delay?

- Incoherent objectives for the model
  - Composition Not defined
  - Objectives expanding constantly

- Solution? Build an early version zero to test, and continue expanding/improving.

What’s next?

V 0.1.1/2/3

- Add Distant Supervision Data to training/FT-ing
- Maybe improved FT-ing methods.

Also need a testing set! - Partly using Matskraft, partly something more (where even matskraft failed)?

Current state

- Data Extraction scripts ready
- Data cleaning or training data generation scripts ready
- FT-ing scripts ready
- Complete pipeline ready

- Need PII’s list from Kausik
- Run Preprocessing pipeline
- Fine-tune the model

Ideas/Features Board

- Handle Testing conditions/Preparation conditions “Descriptions”
- Expand standard forms like EUROFER steel into its constituents even if not present in the current research paper extracting from!
- Maybe add multihop-referencing.

Another idea - maybe for next week?

- a RAG? Can it work?

Accuracies

- Gpt-4o-mini: 40.43%
- Claude-3.5-Haiku: ~50%
- Gemini-2.0-Flash lite: 47.87%
- Gemini-1.5-Flash: 34.78%

Architectures to try!!

Complete paper → Dhruvil’s text extraction (extraction) → LLaMaT (Entity Linkage)

Complete paper → SmolDocling (extraction) → LLaMaT (Entity Linkage)

Complete paper → SmolDocling (extraction) → Any Classical NLP technique? (Entity Linkage)

By Thursday EOD

Converting it from simple composition to list of tuples structure. - by eod today

Run again on the other llms and then run for LLaMat.

Dhruvils extraction might take me some time, so next tuesday meeting!