Northwestern

Cisco Document Intelligence Final Presentation

MLDS Team: Andrew, Kelvin, Xin, Xingyu, Ziqiao

Agenda

Problem Definition

Problem Statement:

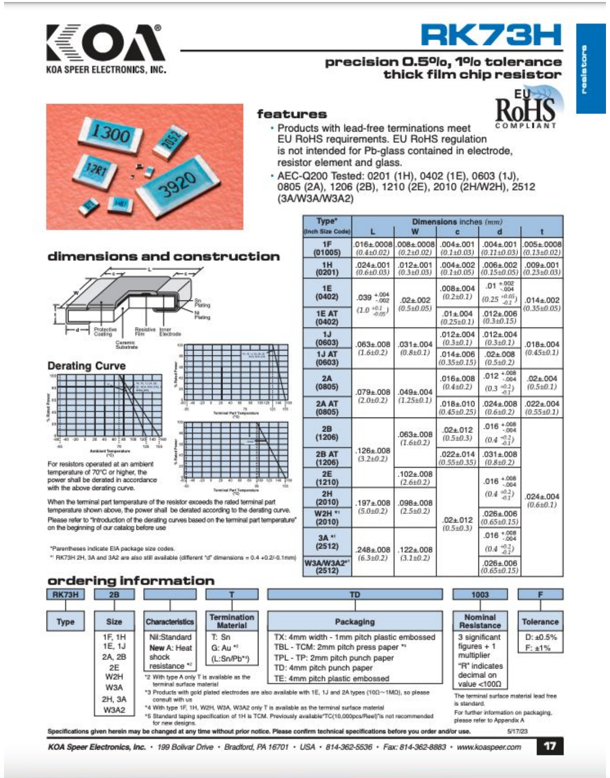
- Cisco supplier product specification files are unstructured
- Product attributes are not consistently formatted or organized within documents
- Current data extraction process is extremely manual and time consuming

Objective:

- Establish an automated data extraction pipeline to collect the information needed from unstructured document sources
- Achieve high accuracy and adaptability to different types of files in the extraction process
- Build an user-friendly UI that non-technical people can interact with
- NOTE: this project will focus on key attribute extraction text-rich PDF files obtained from different suppliers

Example PDF Documents

Document link: ***RK73H.pdf***


KOA SPEER ELECTRONICS, INC.

RK73H

precision 0.5%, 1% tolerance thick film chip resistor

EU RoHS COMPLIANT

resistors

features

- Products with lead-free terminations meet EU RoHS requirements. EU RoHS regulation is not intended for Pb-glass contained in electrode, resistor element and glass.
- AEC-Q200 Tested: 0201 (1H), 0402 (1E), 0603 (1J), 0805 (2A), 1206 (2B), 1210 (2E), 2010 (2H/W2H), 2512 (3A/W3A/W3A2)

dimensions and construction

Derating Curve

For resistors operated at an ambient temperature of 70°C or higher, the power shall be derated in accordance with the above derating curve.

When the terminal part temperature of the resistor exceeds the rated terminal part temperature shown above, the power shall be derated according to the derating curve. Please refer to "Introduction of the derating curves based on the terminal part temperature" on the beginning of our catalog before use

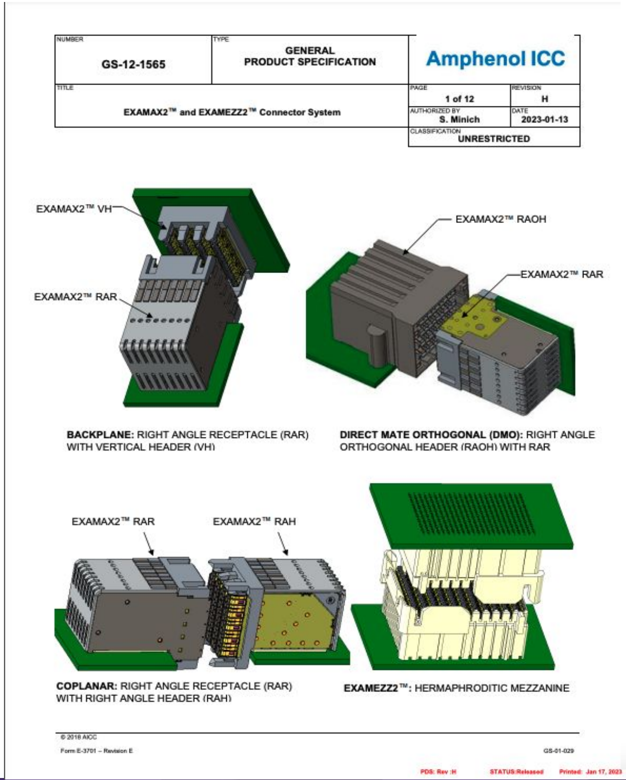
ordering information

RK73H	2B		T	TD	1003	F
Type	Size	Characteristics	Termination Material	Packaging	Nominal Resistance	Tolerance
	1F, 1H, 1E, 1J, 2A, 2B, 2E, W2H, W3A, 2H, 3A, W3A2	Nil:Standard, New A: Heat shock resistance	T: Sn, G: Au, (L:Sn/Pb)	TX: 4mm width - 1mm pitch plastic embossed; TBL - TCM: 2mm pitch press paper; TPL - TP: 2mm pitch punch paper; TD: 4mm pitch punch paper; TE: 4mm pitch plastic embossed	3 significant figures + 1 multiplier "R" indicates decimal on value <100Ω	D: ±0.5%, F: ±1%

Specifications given herein may be changed at any time without prior notice. Please confirm technical specifications before you order and/or use. 5/17/23

KOA Speer Electronics, Inc. • 199 Bolivar Drive • Bradford, PA 16701 • USA • 814-362-5536 • Fax: 814-362-8883 • www.koaspeer.com 17

Document link: Gs-12-1565.pdf


NUMBER	TYPE	
GS-12-1565	GENERAL PRODUCT SPECIFICATION	Amphenol ICC
TITLE: EXAMAX2™ and EXAMEZZ2™ Connector System	PAGE: 1 of 12	REVISION: H
	AUTHORIZED BY: S. Minich	DATE: 2023-01-13
	CLASSIFICATION: UNRESTRICTED	

BACKPLANE: RIGHT ANGLE RECEPTACLE (RAR) WITH VERTICAL HEADER (VH)

DIRECT MATE ORTHOGONAL (DMO): RIGHT ANGLE ORTHOGONAL HEADER (RAOH) WITH RAR

COPLANAR: RIGHT ANGLE RECEPTACLE (RAR) WITH RIGHT ANGLE HEADER (RAH)

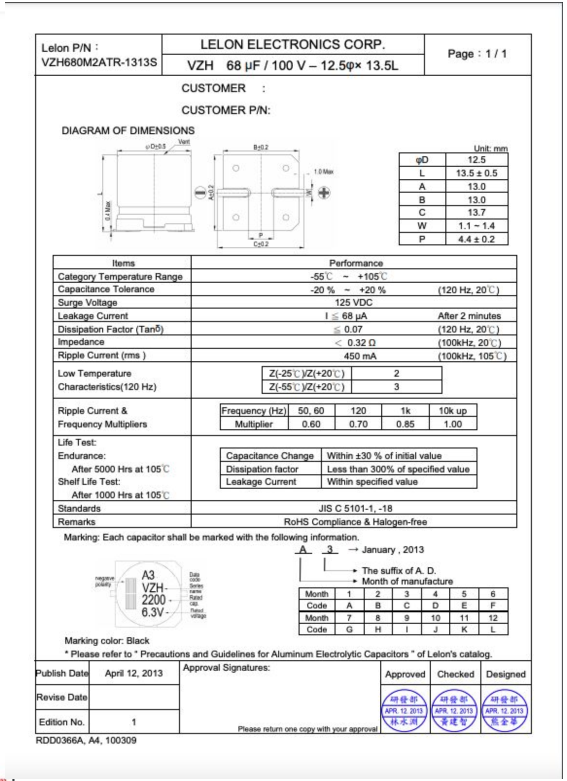
EXAMEZZ2™: HERMAPHRODITIC MEZZANINE

© 2018 AICC

Form E-3701 – Revision E GS-01-029

PDS: Rev :H STATUS:Released Printed: Jan 17, 2023

Document link: ***Vzh680m2atr-1313s.pdf***


Lelon P/N : VZH680M2ATR-1313S	LELON ELECTRONICS CORP. VZH 68 μF / 100 V – 12.5φ× 13.5L	Page : 1 / 1

CUSTOMER :

CUSTOMER P/N:

DIAGRAM OF DIMENSIONS

Unit: mm

φD	12.5
L	13.5 ± 0.5
A	13.0
B	13.0
C	13.7
W	1.1 ~ 1.4
P	4.4 ± 0.2

Items	Performance	
Category Temperature Range	-55℃ ~ +105℃	
Capacitance Tolerance	-20 % ~ +20 %	(120 Hz, 20℃)
Surge Voltage	125 VDC	
Leakage Current	I ≦ 68 μA	After 2 minutes
Dissipation Factor (Tanδ)	≦ 0.07	(120 Hz, 20℃)
Impedance	＜ 0.32 Ω	(100kHz, 20℃)
Ripple Current (rms)	450 mA	(100kHz, 105℃)
Low Temperature Characteristics(120 Hz)	Z(-25℃)/Z(+20℃): 2; Z(-55℃)/Z(+20℃): 3	
Ripple Current & Frequency Multipliers	Frequency (Hz): 50, 60 / 120 / 1k / 10k up; Multiplier: 0.60 / 0.70 / 0.85 / 1.00	
Life Test: Endurance: After 5000 Hrs at 105℃; Shelf Life Test: After 1000 Hrs at 105℃	Capacitance Change: Within ±30 % of initial value; Dissipation factor: Less than 300% of specified value; Leakage Current: Within specified value	
Standards	JIS C 5101-1, -18	
Remarks	RoHS Compliance & Halogen-free	

Marking: Each capacitor shall be marked with the following information.

A 3 → January , 2013

The suffix of A. D.

Month of manufacture

Month	1	2	3	4	5	6
Code	A	B	C	D	E	F
Month	7	8	9	10	11	12
Code	G	H	I	J	K	L

Marking color: Black

* Please refer to " Precautions and Guidelines for Aluminum Electrolytic Capacitors " of Lelon's catalog.

Publish Date	April 12, 2013	Approval Signatures:	Approved	Checked	Designed
Revise Date					
Edition No.	1	Please return one copy with your approval			

RDD0366A, A4, 100309

Data Overview

1. **Attributes have been tested (10 in total):**

   Extracted attributes will depend on product type. Testing was done with the attributes below.

   a. Supplier Name, Product Type, Dimensions, Orientation if any, Current Rating, Voltage, Frequency, Impedance, Capacitance, Temperatures

2. **Selected sample PDF documents (9 in total):**

   All selected products are commodity type for confidentiality reasons and for experimentation purposes

   a. **Capacitors:**
      i. ***KGM15CR51E106K-DATA.pdf*** ***Vzh680m2atr-1313s.pdf***
   b. **Molex:**
      i. 2031430001-PS.pdf
   c. **Amphenol:**
      i. Ssio_mini_cooledge_0_60mm.pdf Gs-12-1457.pdf Gs-12-1565.pdf
   d. **Resistance:**
      i. PE_Lseries.pdf ***RK73H.pdf***
   e. **App Processor:**
      i. ***IMX8MMIEC.pdf***

Note: we selected 9 PDF documents with a diverse range of vendors, products, and document structures. PDFs highlighted in bold and italic are the ones with ground truth values and thus are mainly used for evaluation purposes.

Solution Methodology Overview

- How would we have solved this problem traditionally?
  - A **rule-based/pattern matching approach** (such as RegEx); however, this would have had to be tailored to a specific manufacturer and document type
  - **Machine Learning** based methods (such as SVM, CNN or LSTM) could have provided some generalizability; however, it would require a massive exercise in generating a training data with accurate ground truths

- Why is Generative AI recommended for this use-case?
  - Recent breakthroughs in NLP, particularly with transformer-based models like GPT-4, have significantly **improved the understanding and generation of human-like text**
  - These models have an **enhanced ability to understand context**, which is crucial in accurately extracting relevant information from documents.

Solution Methodology Overview (cont.)

- **What are the advantages/disadvantages of GenAI ?**
  - **Significant Advantages**:
    - *Adaptability and Generalization*: GenAI models are not as rigid as rule-based systems and can generalize better than traditional machine learning models
    - *Contextual Understanding*: GenAI models are better at understanding context and can understand nuances in language and structure that rule-based systems might miss
    - *Multimodal*: new GenAI models are multimodal meaning that they can take in both text and images
  - **Significant Disadvantages**:
    - *Complexity and Resource Intensity*: GenAI models are complex and often require substantial computational resources.
    - *Interpretability and Explainability*: GenAI models are often described as "black boxes" due to their lack of interpretability. Understanding why a model made a particular decision can be challenging, which is problematic in scenarios where transparency is crucial.

Solution Workflow

pdf

Product Spec Documents

→

PDF Extraction Module

Takes a photo of the document and then uses optical character recognition to turn the photo into text

→

TXT

Text-based document

→

LLM (w/ RAG)

Interface with LLMs via "prompts" or requests for action

→

Meaningful Product Attributes

Attributes can be stored in a database, displayed on a website, summarized in a table, etc.

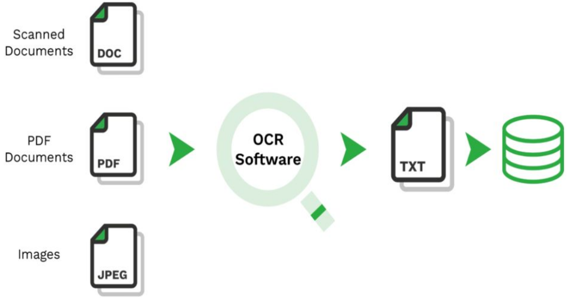
PDF Extraction Module: PDF Pre-Processing

PDF Extraction Methods Overview

PDF extraction is the process of retrieving text, images, data, and metadata from PDF documents using specialized software or techniques, enabling the conversion of content from a non-editable format into a usable, editable format for various applications.

OCR (Optical Character Recognition)-Tesseract

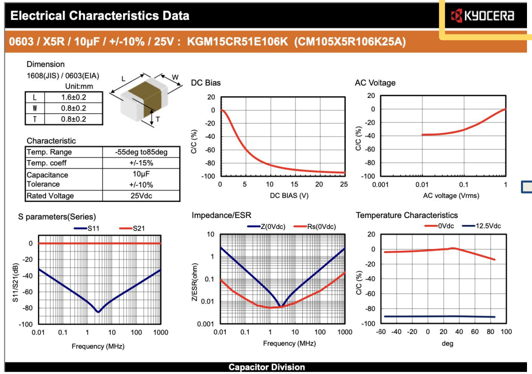
Extracts printed or written text from images. It was originally developed by Hewlett-Packard, and development was later taken over by Google.



Use **Pdf2Image** first convert pdf files to images, then use OCR to extract texts from images.

Input pdf or image

"Image word"



Output a single column of texts

able to extract

```
KYOCERA

Dimension
1608(JIS) / 0603(EIA) L ~ w ;
Unitmm _ a DC Bias AC Voltage
20 20
0 0
Nut -20 _ -20
x. = 40
Characteristic 0 40 o0
Temp. Range -55deg to85deg GS -60 0 -60
Temp. coeff +/-15% -80 -80
Capacitance 10uF -100 es errors -100
Tolerance +/-10% 0 5 10 15 20 25 0.001 0.01 0.1 1
Rated Voltage 25Vdc DC BIAS (V) AC voltage (Vrms)
S parameters(Series) Impedance/ESR Temperature Characteristics
-_-S11 am S71 -=-Z(0Vdc) ==Rs(0Vdc) am-0Vdo me 12.5Vde
10 20
0
0
1
-20
ao E -20
= -40 s iS
a Sf 0.1 XS -40
g o o
~ 60 Wu S -60
" N 0.01
-80 -80
-100 0.001 -100
0.01 0.1 1 10 100 1000 0.01 0.1 1 10 100 1000 -60 -40 -20 0 20 40 60 80 100
Frequency (MHz) deg

Frequency (MHz)

Capacitor Division
```

PDF Pre-Processing: PDF Extraction Methods

Technique/Model	Pros	Cons
pdf2image + OCR-Tesseract	● Open-source and free (can be used locally) ● Good accuracy and versatility ● Can read "image words"	● Requires image preprocessing to achieve optimal results ● Outputs a single column of text, does not preserve the structure
Pdfminer/Pdf Plumber	● Simple and robust; preserves control characters such as "\n"	● Cannot read the "image words" such as vendor name
PyPDF (PyPDF2& PDFLoader)	● PDFLoader has more compatibility with LangChain and so does PyPDF2 with LLaMa	● Some signs cannot be read correctly. E.g. Read °C as oC ● Cannot read "image words"
GPT4.0-image reading	● Good accuracy and versatility ● Can read "image words"	● No accessible API, not free

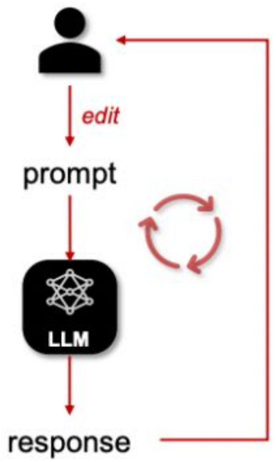
Large Language Models (LLM)

What is a Large Language Model?

- A large language model (LLM) is a type of artificial intelligence that learned on massively large data sets to understand, summarize, generate and predict new content.

How to interact with a Large Language Model?

- Ask clear and specific questions with possible context information
- Follow up on previous questions to improve the answer or ask for new related questions
- **Example use case:** use prompts for question answering, creative tasks like writing stories, poems, or even generating ideas, and more



Large Language Models (LLM)

Open-source vs Closed-source

- **Open-Source Models:** publicly accessible and can be modified or distributed by anyone, can be used without restrictions. Examples: Mistral-7B, Mosaic-mpt7b.
    a. **Pros:** Free to use, customizable, community-driven improvements.
    b. **Cons:** Less support, can require significant resources to train and deploy, potentially lower performance compared to state-of-the-art closed-source models.
- **Closed-Source Models:** These are proprietary models developed by organizations and are not publicly accessible for modification. Example: GPT-3.5, GPT-4.0.
    a. **Pros:** Often more advanced, better performance, reliable support and updates, easier to implement.
    b. **Cons:** Costs (licensing or usage fees), less customizable, reliance on a third-party provider.

Cost Estimates

- **Open-Source Models:** Mostly free, but include potential costs for training (compute resources), maintenance, and integration.
- **Closed-Source Models:** Typically involve licensing fees or pay-per-use models. Provide general cost estimates (if available) or ways to access cost information (e.g., contacting providers for quotes).

Prompt Engineering - The Basics

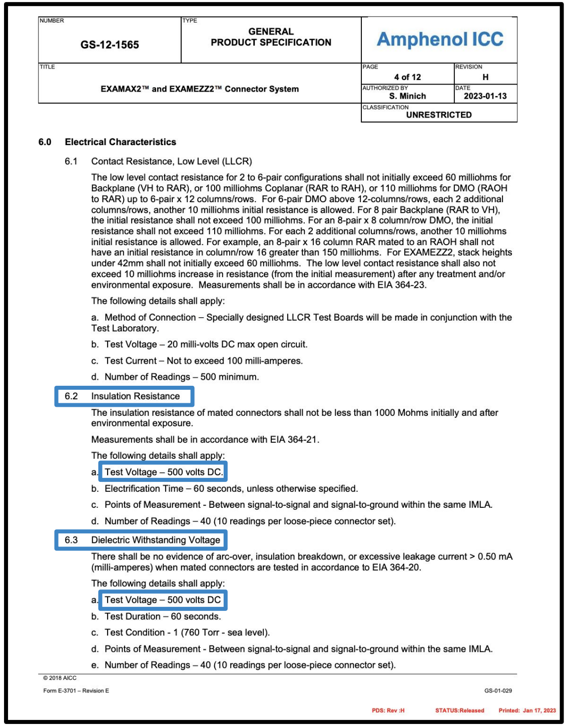
Prompt engineering is the practice of **carefully crafting and optimizing input text** to effectively guide an AI model to produce desired outputs. Benefits of Prompt Engineering include:

- **Enhanced Accuracy**:
    - Modify the prompt to include detailed explanations for each attribute
    - Guide the LLM to provide comprehensive answers that cover all relevant aspects of an attribute but avoid redundancy
    - Rather than asking about all attributes in a single prompt, call the LLM individually for each attribute
- **Avoid Hallucination**:
    - Include a system message in the prompt instructing the LLM to respond with “I don’t know” when it is unsure
- **Correct format**:
    - Provide instruction to the LLM about JSON format and give it an output example. Write a nested dictionary if there are multiple values under one attribute

In summary, revise prompts for improved LLM performance, while considering the effort involved in prompt engineering and factoring in the token costs associated with longer prompts.

Example Attribute Extraction

Extraction Method: Pytesseract
Model: GPT-4
Data: gs-12-1565.pdf



NUMBER	TYPE	
GS-12-1565	GENERAL PRODUCT SPECIFICATION	Amphenol ICC

TITLE	PAGE	REVISION
EXAMAX2™ and EXAMEZZ2™ Connector System	4 of 12	H
	AUTHORIZED BY: S. Minich	DATE: 2023-01-13
	CLASSIFICATION: UNRESTRICTED	

6.0 Electrical Characteristics

6.1 Contact Resistance, Low Level (LLCR)

The low level contact resistance for 2 to 6-pair configurations shall not initially exceed 60 milliohms for Backplane (VH to RAR), or 100 milliohms Coplanar (RAR to RAH), or 110 milliohms for DMO (RAOH to RAR) up to 6-pair x 12 columns/rows. For 6-pair DMO above 12-columns/rows, each 2 additional columns/rows, another 10 milliohms initial resistance is allowed. For 8 pair Backplane (RAR to VH), the initial resistance shall not exceed 100 milliohms. For an 8-pair x 8 column/row DMO, the initial resistance shall not exceed 110 milliohms. For each 2 additional columns/rows, another 10 milliohms initial resistance is allowed. For example, an 8-pair x 16 column RAR mated to an RAOH shall not have an initial resistance in column/row 16 greater than 150 milliohms. For EXAMEZZ2, stack heights under 42mm shall not initially exceed 60 milliohms. The low level contact resistance shall also not exceed 10 milliohms increase in resistance (from the initial measurement) after any treatment and/or environmental exposure. Measurements shall be in accordance with EIA 364-23.

The following details shall apply:

a. Method of Connection – Specially designed LLCR Test Boards will be made in conjunction with the Test Laboratory.

b. Test Voltage – 20 milli-volts DC max open circuit.

c. Test Current – Not to exceed 100 milli-amperes.

d. Number of Readings – 500 minimum.

6.2 Insulation Resistance

The insulation resistance of mated connectors shall not be less than 1000 Mohms initially and after environmental exposure.

Measurements shall be in accordance with EIA 364-21.

The following details shall apply:

a. Test Voltage – 500 volts DC.

b. Electrification Time – 60 seconds, unless otherwise specified.

c. Points of Measurement - Between signal-to-signal and signal-to-ground within the same IMLA.

d. Number of Readings – 40 (10 readings per loose-piece connector set).

6.3 Dielectric Withstanding Voltage

There shall be no evidence of arc-over, insulation breakdown, or excessive leakage current > 0.50 mA (milli-amperes) when mated connectors are tested in accordance to EIA 364-20.

The following details shall apply:

a. Test Voltage – 500 volts DC

b. Test Duration – 60 seconds.

c. Test Condition - 1 (760 Torr - sea level).

d. Points of Measurement - Between signal-to-signal and signal-to-ground within the same IMLA.

e. Number of Readings – 40 (10 readings per loose-piece connector set).

© 2018 AICC

Form E-3701 – Revision E GS-01-029

PDS: Rev :H STATUS:Released Printed: Jan 17, 2023

Example 2: Voltage Extraction

Voltage:
500 volts DC; 6.3 Dielectric Withstanding Voltage,
500 volts DC; 6.2 Insulation Resistance

Example Attribute Extraction

Extraction Method: Pytesseract
Model: GPT-4
Data: gs-12-1565.pdf

9.2 Inspection Conditions

Unless otherwise specified herein, all inspections shall be performed under the following ar
conditions:

a. Temperature: 25 +/- 5 deg C

b. Relative Humidity: 30% to 60%

c. Barometric Pressure: Local ambient

8.3 High Temperature Life –EIA 364-17

a. Method A

b. Test Temperature - 85 deg C

c. Test Duration - 500 hours

8.1 Thermal Shock – EIA 364-32.

a. Number of Cycles - 5

b. Temperature Range Between -55 and +85 deg C

c. Time at Each Temperature - 30 minutes minimum

d. Transfer Time - 30 seconds, maximum

Example 1: Temperature Extraction

25+/-5 degC; 9.2 Inspection Conditions
-55 and +85 deg C; 8.1 Thermal Shock
85 deg C, 8.3 High Temperature Life

Prompt Engineering - Basic Prompt

Sample Prompts	Output Example	Observations
"I'm looking for information in this document. Here is the information I'm looking for Supplier Name, Product Name, Product Type, Dimensions, Orientation, Current Rating, Voltage, Frequency, Impedance / Capacitance, Temperatures"	By GPT-4: • The supplier name is Amphenol ICC. The product names are EXAMAX2™ and EXAMEZZ2™ Connector System. The product type is Mezzanine connectors. • The document does not provide specific information on dimensions and orientation. • The operating current rating can be found in the AICC Power Design Guide GS-20-0456. • The agency voltage rating is less than 30 VAC RMS / DC and the non-agency voltage rating is 150 Vac RMS or Vpc Maximum. • The document does not provide information on frequency and impedance/capacitance. The operating temperature range is -55 to 85 °C.	• Output **texts** with a list of attributes and results. • A slash in the attribute list was not conducive to the final output structure. Sometimes it would separate the two attributes sometimes keep them together.

Prompt Engineering - Final Prompt

<table>
<tr><th>Sample Prompts</th><th>Output Example</th><th>Observations</th></tr>
<tr><td>
System message: Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.
Prompt: Please extract and compile all information related to the following attributes from the document, avoiding any duplication. For each attribute, include the specific source or context for each unique value. Present the findings of each attribute in a single dictionary entry.
Exclude any redundant information.
(a) Supplier Name
...
(i) Temperatures(temperatures under any conditions)

The answer should be in the following format:
{
"attribute": "Supplier Name",
"values": [
{"value": "unique value 1", "source": "context or source"},
{"value": "unique value 2", "source": "context or source"}
],
"summary": "Provide a general summary or observation about the extracted information, focusing on unique aspects."
},
{
"attribute": "Product Type",
...
},for dimensions, orientation and so on.
</td><td>
By GPT-4:
{
"attribute": "Temperatures",
"values":
[{"value": "25 +/-5 degC", "source": "9.2 Inspection Conditions"},
{"value": "-55 and +85 deg C", "source": "8.1 Thermal Shock"},
{"value": "85 deg C", "source": "8.3 High Temperature Life"}],
"summary": "The document provides information on temperatures under different conditions.”
},
{
"attribute": "Voltage",
"values":
[{"value": "500 volts DC", "source": "6.3 Dielectric Withstanding Voltage"},
{"value": "500 volts DC", "source": "6.2 Insulation Resistance"}],
"summary": "...”
}
</td><td>

Output JSON with nested dictionaries of attribute names, multiple attribute values and source.
Both GPT 4 and GPT 3.5 are able to output in the instructed format, and provide reference (e.g. explanation, section number or section title)

</td></tr>
</table>

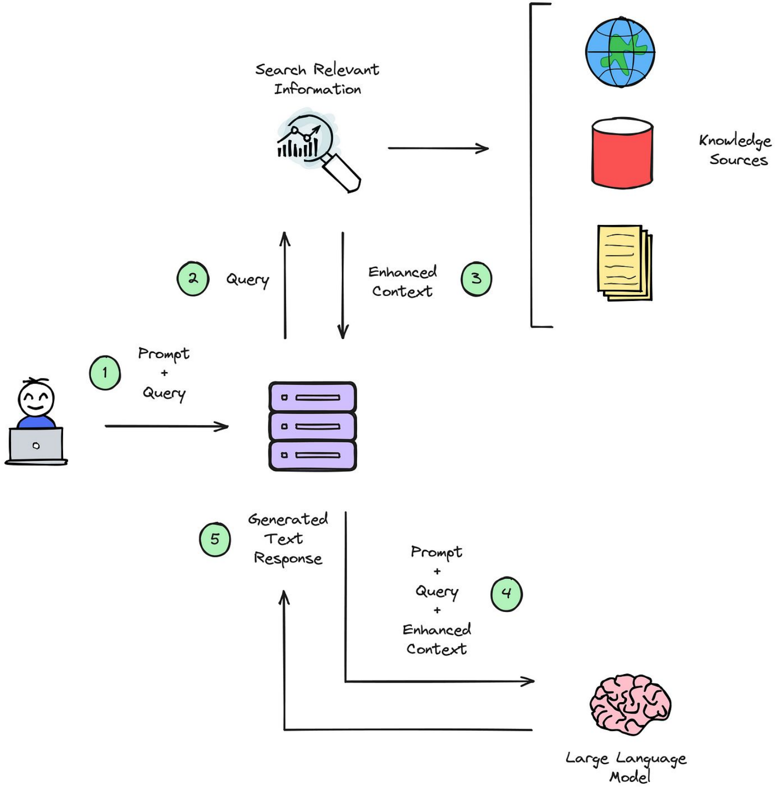
Multi-Prompt vs Single-Prompt

Prompt	Definition	Pros	Cons
Multi-Prompt	Individual prompt for each attribute and call LLM API separately.	• More precise, relevant and comprehensive answers • Error in one response doesn't affect the extraction of other attributes. • Allows for tailored prompts for each attribute, thus handling complex attributes more effectively.	• More API calls, more tokens, higher cost • Sequential processing can be slower
Single-Prompt	One prompt for all attributes within a document and call API once.	• Fewer API calls, fewer tokens, lower cost • Faster especially for less complex document	• Less precise responses • Error in the response can affect the extraction of all attributes. • Limited customization

Enhancing LLM with Retrieval Augmented Generation

Steps for Retrieval Augmented Generation (RAG):

1. **Prompt + Query**: The user starts the process by entering a query that reflects their question or topic of interest
2. **Query**: For each query, the system conducts a similarity search against the knowledge sources to find relevant information
3. **Enhanced Context**: Contextual information closely matching the query is extracted from the knowledge sources to serve as reference information
4. **Prompt + Query + Enhanced Text**: This extracted context, along with the user's original prompt and query, is synthesized to be inputed for the Large Language Model (LLM)
5. **Generated Text Response**: The LLM processes the integrated input to produce a precise and contextually relevant answer to the user's query



Why we chose RAG based approach

With RAG:

- Can be paired with specific domain databases to provide responses in specialized fields.
- Can split longer texts into chunks.
- Can provide the source text.
- Can introduce latency.
- There's no guarantee that all information will be preserved in the result (potential information loss).

Without RAG:

- Works well with shorter PDFs (e.g., 1 page).
- Does not introduce latency.
- Can't work with longer PDFs (token limit issue).
- Can't locate the source text.
- Tends to hallucinate more

We finally choose to use RAG because:

- It allows pairing with specific domain databases for specialized responses.
- It handles longer and complex texts by splitting them into chunks.
- It provides the source text for transparency, even though there might be a latency issue.
- It offers the flexibility to work with a variety of information sources.

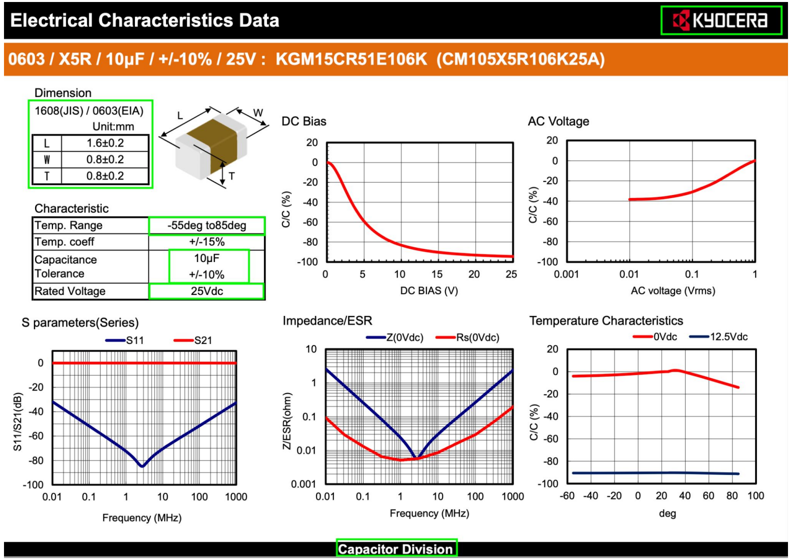
Method Comparison - LLM

Models	Pros	Cons	Decision
GPT3.5	● More accurate than open-source models	● Less accurate and advanced than GPT-4.0	We use as final model for this project considering the balance between cost and performance
GPT4.0	● More accurate than GPT 3.5 on complex tasks	● More expensive than GPT 3.5	Recommend to use for pdf reading if budget allows
GPT4.0-turbo	● Can directly read image through API	● New model and high cost	Recommend to use if needs to deal with highly unstructured image information
Mistral-7B	● Completely free and secure	● Unable to detect some information of the attributes	Not suggested
Mosaic-mpt7b (instruct)	● Completely free and secure and accepts very long token inputs	● Hard to convince the model to follow the exact instruction ● Bad performance	Not suggested

Result Comparison with Ground Truth

- Short Documents with clear tables

Extraction Method: Pytesseract
Model: GPT 3.5
Data: KGM15CR51E106K-DATA.pdf



Example 1:

The GPT- 3.5 model is able to correctly locate relevant information for attributes applicable to this specific document, **achieving 100% accuracy** with no manual proofing needed for approval.

PDF Name	Answer Type	Supplier Name	Product Type	Dimension	Voltage	Capacitance	Temperature
KGM15CR51E106K-DATA.pdf	Model Output	KYOCERA	Capacitor	1608(JIS) / 0603(EIA) L ~ w	25Vdc	10uF	-55deg to85deg
	Ground True Value	KYOCERA	Capacitor	1608(JIS) / 0603(EIA) L ~ w	25V	10 uF	-55C to 85 C

Result Comparison with Ground Truth

- Short Documents with clear tables

Extraction Method: Pytesseract
Model: GPT 3.5
Data: *Vzh680m2atr-1313s.pdf*

Example 2:

The result is still **almost 100% accurate** except for one parameter with extra width

PDF Name	Answer Type	Supplier Name	Product Type	Dimension
Vzh680m2atr-1313s.pdf	Model Output	LELON ELECTRONICS	Capacitor	Length': '12.5 mm', 'Width': '10 mm', 'Height': '13.5 mm'
	Ground True Value	LELON ELECTRONICS	Capacitor	Length:12.5 mm, Height: 13.5mm

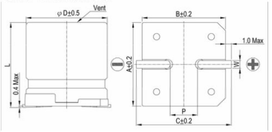
PDF Name	Answer Type	Voltage	Impedance	Capacitance	Temperature
Vzh680m2atr-1313s.pdf	Model Output	100 V	< 0.320 (100kHz, 20°C)	68 uF	-55°C ~ +105°C
	Ground True Value	2A (100 V)	< 0.320 (100kHz, 20°C)	68 uF	-55C to 105C

Lelon P/N：VZH680M2ATR-1313S	LELON ELECTRONICS CORP. VZH 68 μF / 100 V – 12.5φ× 13.5L	Page：1 / 1

CUSTOMER :

CUSTOMER P/N:

DIAGRAM OF DIMENSIONS



Unit: mm

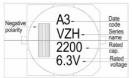
φD	12.5
L	13.5 ± 0.5
A	13.0
B	13.0
C	13.7
W	1.1 ~ 1.4
P	4.4 ± 0.2

Items	Performance	
Category Temperature Range	-55℃ ~ +105℃	
Capacitance Tolerance	-20 % ~ +20 %	(120 Hz, 20℃)
Surge Voltage	125 VDC	
Leakage Current	I ≦ 68 μA	After 2 minutes
Dissipation Factor (Tanδ)	≦ 0.07	(120 Hz, 20℃)
Impedance	< 0.32 Ω	(100kHz, 20℃)
Ripple Current (rms)	450 mA	(100kHz, 105℃)
Low Temperature Characteristics(120 Hz)	Z(-25℃)/Z(+20℃) 2; Z(-55℃)/Z(+20℃) 3	
Ripple Current & Frequency Multipliers	Frequency (Hz): 50, 60 / 120 / 1k / 10k up; Multiplier: 0.60 / 0.70 / 0.85 / 1.00	
Life Test: Endurance: After 5000 Hrs at 105℃; Shelf Life Test: After 1000 Hrs at 105℃	Capacitance Change: Within ±30 % of initial value; Dissipation factor: Less than 300% of specified value; Leakage Current: Within specified value	
Standards	JIS C 5101-1, -18	
Remarks	RoHS Compliance & Halogen-free	

Marking: Each capacitor shall be marked with the following information.

A 3 → January , 2013

→ The suffix of A. D.
→ Month of manufacture



Month	1	2	3	4	5	6
Code	A	B	C	D	E	F
Month	7	8	9	10	11	12
Code	G	H	I	J	K	L

Marking color: Black

* Please refer to " Precautions and Guidelines for Aluminum Electrolytic Capacitors " of Lelon's catalog.

Publish Date	April 12, 2013	Approval Signatures:	Approved	Checked	Designed
Revise Date			研發部 APR. 12. 2013 林水淵	研發部 APR. 12. 2013 黃建智	研發部 APR. 12. 2013 熊金華
Edition No.	1	Please return one copy with your approval			

RDD0366A, A4, 100309

Result Comparison with Ground Truth

- Long Documents with complex tables

Extraction Method: Pytesseract
Model: GPT 3.5
Data: RK73H.pdf IMX8MMIEC.pdf

Example 3 (2 pages):

PDF Name	Answer Type	Supplier Name	Product Type	Dimension
RK73H.pdf	Model Output Ground True Value	Not found KOA Speer Electronics, Inc	Resistor Resistor	Length: 0.016 inches (0.4 mm), Width: 0.008 inches 1E (0402); 0.04 in X 0.02 in

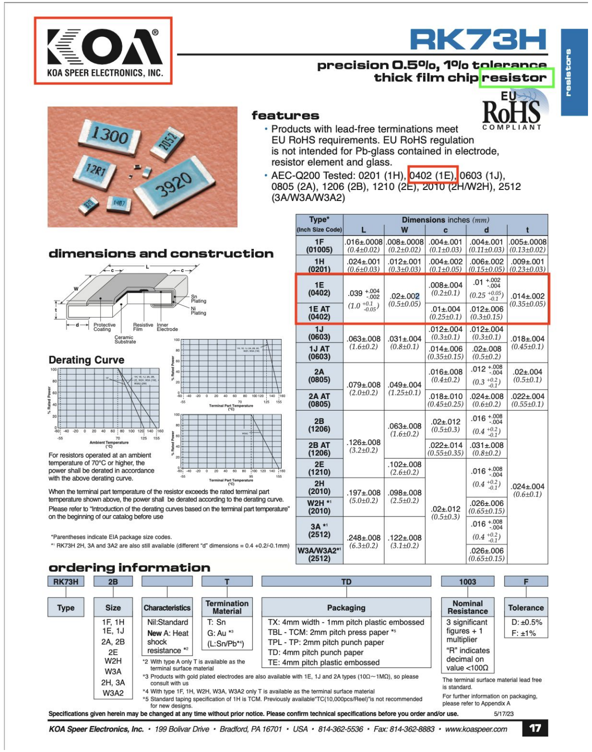
Example 4 (96 pages):

PDF Name	Answer Type	Supplier Name	Product Type	Dimension
IMX8MMIEC.pdf	Model Output Ground True Value	NXP Semiconductors NXP Semiconductors	Not found (App Processor) i.MX 8M M	14 x 14 mm Package Length (mm):14 mm, Package Width (mm):14 mr

Without using the image API, PDFs which are extremely long and have complex formattings and tables, the model's accuracy is negatively impacted.

In these cases, manual check is required to validate results.

Example 3 document (Page 1/2)



KOA SPEER ELECTRONICS, INC.

RK73H

precision 0.5%, 1% tolerance
thick film chip resistor

resistors

EU RoHS COMPLIANT

features

- Products with lead-free terminations meet EU RoHS requirements. EU RoHS regulation is not intended for Pb-glass contained in electrode, resistor element and glass.
- AEC-Q200 Tested: 0201 (1H), 0402 (1E), 0603 (1J), 0805 (2A), 1206 (2B), 1210 (2E), 2010 (2H/W2H), 2512 (3A/W3A/W3A2)

dimensions and construction

Derating Curve

For resistors operated at an ambient temperature of 70°C or higher, the power shall be derated in accordance with the above derating curve.

When the terminal part temperature of the resistor exceeds the rated terminal part temperature shown above, the power shall be derated according to the derating curve.

Please refer to "Introduction of the derating curves based on the terminal part temperature" on the beginning of our catalog before use

*Parentheses indicate EIA package size codes.
*1 RK73H 2H, 3A and 3A2 are also still available (different "d" dimensions = 0.4 +0.2/-0.1mm)

Type* (Inch Size Code)	Dimensions inches (mm) L	W	c	d	t
1F (01005)	.016±.0008 (0.4±0.02)	.008±.0008 (0.2±0.02)	.004±.001 (0.1±0.03)	.004±.001 (0.11±0.03)	.005±.0008 (0.13±0.02)
1H (0201)	.024±.001 (0.6±0.03)	.012±.001 (0.3±0.03)	.004±.002 (0.1±0.05)	.006±.002 (0.15±0.05)	.009±.001 (0.23±0.03)
1E (0402)	.039 +.004 -.002 (1.0 +0.1 -0.05)	.02±.002 (0.5±0.05)	.008±.004 (0.2±0.1)	.01 +.002 -.004 (0.25 +0.05 -0.1)	.014±.002 (0.35±0.05)
1E AT (0402)			.01±.004 (0.25±0.1)	.012±.006 (0.3±0.15)	
1J (0603)	.063±.008 (1.6±0.2)	.031±.004 (0.8±0.1)	.012±.004 (0.3±0.1)	.012±.004 (0.3±0.1)	.018±.004 (0.45±0.1)
1J AT (0603)			.014±.006 (0.35±0.15)	.02±.008 (0.5±0.2)	
2A (0805)	.079±.008 (2.0±0.2)	.049±.004 (1.25±0.1)	.016±.008 (0.4±0.2)	.012 +.008 -.004 (0.3 +0.2 -0.1)	.02±.004 (0.5±0.1)
2A AT (0805)			.018±.010 (0.45±0.25)	.024±.008 (0.6±0.2)	.022±.004 (0.55±0.1)
2B (1206)	.126±.008 (3.2±0.2)	.063±.008 (1.6±0.2)	.02±.012 (0.5±0.3)	.016 +.008 -.004 (0.4 +0.2 -0.1)	.024±.004 (0.6±0.1)
2B AT (1206)			.022±.014 (0.55±0.35)	.031±.008 (0.8±0.2)	
2E (1210)		.102±.008 (2.6±0.2)	.02±.012 (0.5±0.3)	.016 +.008 -.004 (0.4 +0.2 -0.1)	
2H (2010)	.197±.008 (5.0±0.2)	.098±.008 (2.5±0.2)			
W2H *1 (2010)				.026±.006 (0.65±0.15)	
3A *1 (2512)	.248±.008 (6.3±0.2)	.122±.008 (3.1±0.2)		.016 +.008 -.004 (0.4 +0.2 -0.1)	
W3A/W3A2*1 (2512)				.026±.006 (0.65±0.15)	

ordering information

RK73H	2B		T	TD	1003	F
Type	Size	Characteristics	Termination Material	Packaging	Nominal Resistance	Tolerance
	1F, 1H 1E, 1J 2A, 2B 2E W2H W3A 2H, 3A W3A2	Nil:Standard New A: Heat shock resistance *2	T: Sn G: Au *3 (L:Sn/Pb*4)	TX: 4mm width - 1mm pitch plastic embossed TBL - TCM: 2mm pitch press paper *5 TPL - TP: 2mm pitch punch paper TD: 4mm pitch punch paper TE: 4mm pitch plastic embossed	3 significant figures + 1 multiplier "R" indicates decimal on value <100Ω	D: ±0.5% F: ±1%

*2 With type A only T is available as the terminal surface material
*3 Products with gold plated electrodes are also available with 1E, 1J and 2A types (10Ω~1MΩ), so please consult with us
*4 With type 1F, 1H, W2H, W3A, W3A2 only T is available as the terminal surface material
*5 Standard taping specification of 1H is TCM. Previously available"TC(10,000pcs/Reel)"is not recommended for new designs.

The terminal surface material lead free is standard.

For further information on packaging, please refer to Appendix A

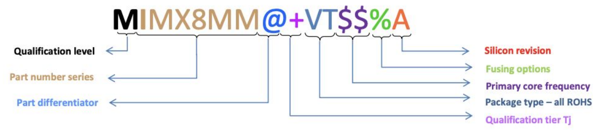
Specifications given herein may be changed at any time without prior notice. Please confirm technical specifications before you order and/or use. 5/17/23

KOA Speer Electronics, Inc. • 199 Bolivar Drive • Bradford, PA 16701 • USA • 814-362-5536 • Fax: 814-362-8883 • www.koaspeer.com 17

Challenge Scenario – Future Results

- **Certain product documents requires “decoding” of part name nomenclature (see picture below)**
  - While testing, the ChatGPT API could only parse text, and therefore could not capture the formatting
  - However, the ChatGPT 4.0 web app is multimodal, and therefore can accept document images
  - NOTE: As of mid November, the OpenAI API can accommodate pictures

- **Multimodal GPT 4.0 still requires specific prompting to parse the data**
  - When asked to decode the picture (to the right), the model was generally able to summarize the tables; however there were small inaccuracies, especially with Package Type
  - When prompted to decode a part number never before seen, the model was able to properly decode the part number and list the attributes



Qualification Level	
Samples	P
Mass Production	M

Part number series	Name
IMX8MM	i.MX 8M Mini

Part differentiator	@
i.MX 8M Mini Quad 4x A53, M4, GPU, VPU	6
i.MX 8M Mini QuadLite 4x A53, M4, GPU	5
i.MX 8M Mini Dual 2x A53, M4, GPU, VPU	4
i.MX 8M Mini DualLite 2x A53, M4, GPU	3
i.MX 8M Mini Solo 1x A53, M4, GPU, VPU	2
i.MX 8M Mini SoloLite 1x A53, M4, GPU	1

Temperature Tj	+
Consumer: 0 to +95ºC	D
Industrial: -40 to 105ºC	C

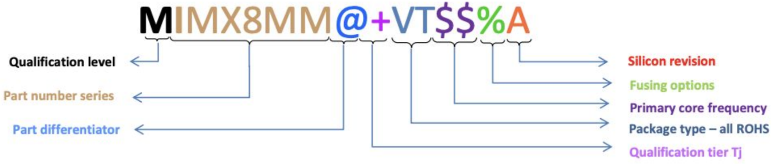
Package Type	ROHS
FCBGA486 14 x 14 mm, 0.5 mm pitch	VT

Frequency	$$
1.8 GHz	LZ
1.6 GHz	KZ

Fusing	%
Default	A
Immersiv3D enabled w/Dolby Atmos	C
Immersiv3D enabled w/Dolby Atmos and DTS	D

Silicon rev	A
Rev A0	A

Challenge Scenario – Future Results



Qualification Level	
Samples	P
Mass Production	M

Part number series	Name
IMX8MM	i.MX 8M Mini

Part differentiator	@
i.MX 8M Mini Quad 4x A53, M4, GPU, VPU	6
i.MX 8M Mini QuadLite 4x A53, M4, GPU	5
i.MX 8M Mini Dual 2x A53, M4, GPU, VPU	4
i.MX 8M Mini DualLite 2x A53, M4, GPU	3
i.MX 8M Mini Solo 1x A53, M4, GPU, VPU	2
i.MX 8M Mini SoloLite 1x A53, M4, GPU	1

Temperature Tj	+
Consumer: 0 to +95ºC	D
Industrial: -40 to 105ºC	C

Package Type	ROHS
FCBGA486 14 x 14 mm, 0.5 mm pitch	VT

Frequency	$$
1.8 GHz	LZ
1.6 GHz	KZ

Fusing	%
Default	A
Immersiv3D enabled w/Dolby Atmos	C
Immersiv3D enabled w/Dolby Atmos and DTS	D

Silicon rev	A
Rev A0	A

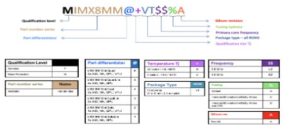


Figure 2. Part number nomenclature—i.MX 8M Mini family of processors

can you decode this

ChatGPT

The image you've uploaded appears to be a diagram explaining the part number nomenclature for a family of processors, specifically the i.MX 8M Mini family. The part number nomenclature is broken down into several components to specify different product options. Here's what each section seems to represent:

1. **Qualification Level**:
   - P: Samples
   - M: Mass Production
2. **Part Number Series**:
   - The name of the part number series is indicated as i.MX8M for the i.MX 8M Mini family of processors.
3. **Part Differentiator**:
   - This is a numerical value that differentiates between the types of processors within the i.MX 8M Mini family, indicating different configurations such as Quad, DualLite, Solo, etc., along with the supported features like GPU, VPU, etc.
4. **Temperature TJ**:
   - Consumer-grade temperature range: 0 to +95°C
   - Industrial-grade temperature range: -40 to +105°C
5. **Package Type**:
   - FCBGA486: This indicates the package type of the processor, which is a 486-ball Flip Chip Ball Grid Array with a certain pitch size.
6. **Silicon Revision**:
   - The letter (e.g., A) indicates the revision of the silicon used in the processor.
7. **Fusing Options**:
   - Default
   - Immersiv3D enabled w/Dolby Atmos
   - Immersiv3D enabled w/Dolby Atmos and DTS
8. **Primary Core Frequency**:
   - Different letters are used to represent the core frequency of the processor (e.g., 1.8 GHz, 1.6 GHz).
9. **Fusing, Silicon Revision, Frequency, and Package Type**:
   - Symbols (@, +, $$, %, A) and letters (VT, LZ, KZ, C, D) are used as codes to represent different options and features for the processors.

The nomenclature is essentially a structured code that allows for precise identification of processor versions and their features. It's a standardized way for manufacturers and customers to communicate about exactly which ver[...]or they are referring to.

Challenge Scenario – Future Results

YA **You**

using the encoding above, what is MIMX8MM5CVTKZAA

ChatGPT

The part number "MIMX8MM5CVTKZAA" can be decoded using the key provided in the image you've uploaded. Let's break it down:

- **MI**: This prefix is not explicitly explained in the diagram, but it typically stands for the manufacturer's code or product family.
- **MX8MM**: This refers to the product family, the i.MX 8M Mini.
- **5**: This digit is the part differentiator which specifies the type of processor. According to the key:
  - 5 indicates an i.MX 8M Mini QuadLite processor, which has 4x A53, M4, GPU cores.
- **C**: This letter indicates the temperature range. C corresponds to Industrial-grade, which is -40 to +105°C.
- **VT**: This code specifies the package type, which in this case is FCBGA486, a 486-ball Flip Chip Ball Grid Array with a 0.5 mm pitch.
- **KZ**: These two letters indicate the primary core frequency. KZ corresponds to a frequency of 1.6 GHz.
- **AA**: Typically, these letters at the end would denote the silicon revision and fusing options or additional features. However, the provided key does not detail what 'AA' stands for. In common practice, this could stand for a default configuration or a first revision without special options.

So, the part number "MIMX8MM5CVTKZAA" refers to an i.MX 8M Mini QuadLite processor with an industrial temperature range, packaged in a FCBGA486 with a 1.6 GHz core frequency, and possibly a default silicon revision and fusing options.

YA **You**

Do it again, but clearly denote: Qualification level, Part number series, Part differentiator, Silicon revision, Fusing options, Primary core frequency, Package type – all ROHS, Qualification tier Tj

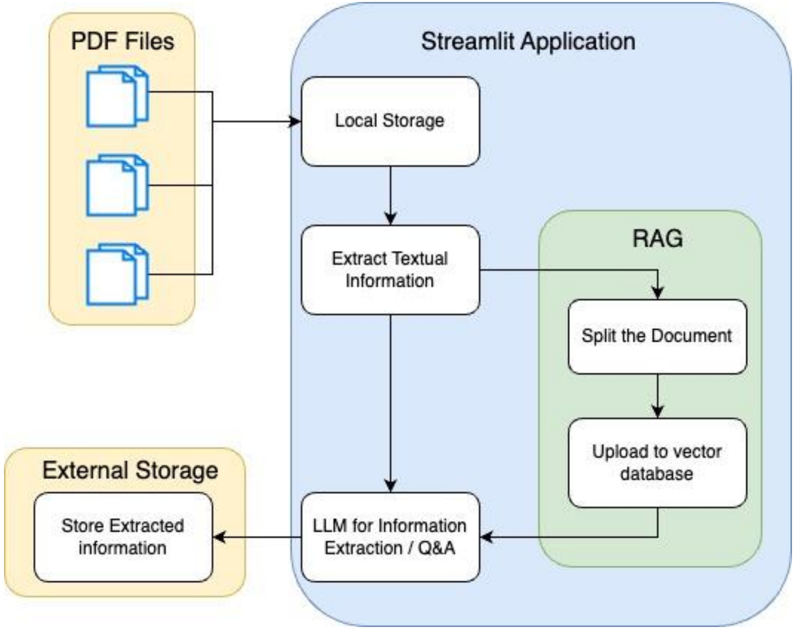
ChatGPT

Certainly, let's decode the part number "MIMX8MM5CVTKZAA" with the specific sections clearly denoted:

1. **Qualification Level**:
   - **M**: Indicates that the part is approved for Mass Production.
2. **Part Number Series**:
   - **IMX8MM**: Refers to the i.MX 8M Mini series.
3. **Part Differentiator**:
   - **5**: Corresponds to the i.MX 8M Mini QuadLite version, with 4x ARM Cortex-A53 cores, 1x Cortex-M4 core, and GPU.
4. **Silicon Revision**:
   - The provided key does not specify how to decode the Silicon Revision from this part number. Normally it is indicated by a letter, but in this part number, the letter for Silicon Revision is not clear.
5. **Fusing Options**:
   - The provided key does not specify how to decode the Fusing Options from this part number. Normally it is indicated by a character, but in this part number, the character for Fusing Options is not clear.
6. **Primary Core Frequency**:
   - **KZ**: This indicates a primary core frequency of 1.6 GHz.
7. **Package Type – all ROHS**:
   - **VT**: Signifies the package type, which is FCBGA486 (14 × 14 mm, 0.5 mm pitch), and it complies with the ROHS standards for restriction of hazardous substances.
8. **Qualification Tier Tj**:
   - **C**: Denotes an industrial qualification tier with a junction temperature (Tj) range of -40 to +105°C.

In summary, the part number "MIMX8MM5CVTKZAA" describes an i.MX 8M Mini QuadLite processor, suitable for mass production, with an industrial temperature range, a 1.6 GHz primary core frequency, and packaged in a ROHS-compliant FCBGA486. The silicon revision and fusing options are not specified with the informat ed.

Demo



Major Pipeline Steps

- **Step 1:** PDF files are inputted and temporarily stored for preprocessing.
- **Step 2:** Textual information is extracted from the PDF files using OCR (tesseract).
- **Step 3:** The RAG (retrieval-augmented generation) module processes the text, splitting the documents into manageable segments for further analysis.
- **Step 4:** These segments are then uploaded to a vector database, where they're indexed for efficient retrieval.
- **Step 5:** A conversational chain utilizes the processed data for information extraction and question-answering (Q&A) tasks.

Demo

Sample LLM Costs

- **LLM Costs are calculated based on “tokens” which are groups of characters, which sometimes align with words, but not always.**
  - **1 token ~= 4 chars in English. 1 token ~= ¾ words. 100 tokens ~= 75 words**
  - Costs can be broken into Input Costs and Output Costs, both of which can be managed
  - Costs vary based on model (more powerful models are more expensive)

Cost for 1K Documents

Model	Context length	Cost for Input	Cost for Output	Total Costs
GPT 3.5-turbo	16k	$ 3.80	$ 1.00	**$ 4.80**
GPT 4.0-turbo	128k	$ 38.00	$ 15.00	**$ 53.00**
GPT 4.0	32k	$ 228.00	$ 60.00	**$ 288.00**

Note: costs assumes average input of 3800 tokens and output of 200 tokens
OpenAI Pricing: https://openai.com/pricing

Next Steps and Recommendations

1. **OpenAI just released picture handling in their new API; based on initial testing, we think this method will be immensely useful for Cisco:**
   a. Product documents that require “decoding” as previously mentioned
   b. Reading and summarizing graphs and line charts
   c. Reading and interpreting data tables (initial tests conversion of data tables to JSON format were successful, though will require prompt tuning)
2. **Human-in-the-loop validation: complex document types and documents of a certain length would use humans for validation while simple, shorter documents would rely solely on the pipeline**
3. **Customized prompt with attribute explanation works better than a generic prompt, but prompt engineering effort should be considered.**
4. **Given the additional model performance and extended context window, we suggest future builds use the GPT4.0-turbo model**

Next Steps and Recommendations (cont.)

- To determine the final prompt of choice, we suggest following our defined “Success Framework” which scores Model Output against Ground Truth across many documents (>30 preferred)
- Given the constantly changing technical environment, we suggest building the final pipeline in a modular fashion by keeping the RAG, Model Itself, File Reading Mechanism, and Web App independent of one another to allow for simple “lift and shifts”

Milestones (Proposed)

1. **Stage 1: Single Document Attribute Extraction (Oct 1st - Oct 15)**
    - Use a variety of methods; results will dictate which method(s) we'll continue forward with
2. **Stage 2: Cross-Product Attribute Extraction from Same Supplier (Oct 16 - Oct 31)**
    - Continue to narrow down on methodology
3. **Mid-Stage Review (Nov 1st)**

We are here

---

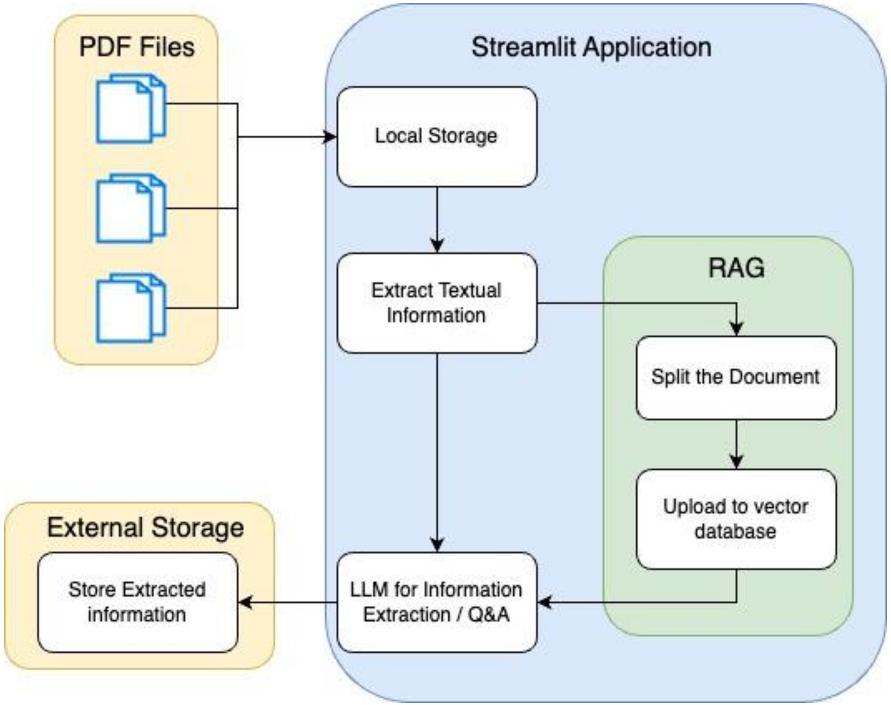
4. **Stage 3: Cross-Supplier and Cross-Product Attribute Extraction (Nov 1st - Nov 15)**
    - Test the extraction process on suppliers and product data, refine the algorithm based on test results
    - **(updated)** Validate the accuracy, consistency, and costs of the final attribute extraction method
    - **(new)** Get approval from the Cisco team on the final architecture
5. **(updated) Stage 4: Chat-Bot, Web Application, and Final Pipeline Development (Nov 16 - Dec 6th)**
6. **Final Review and Presentation (Dec 6th)**
    - Review the entire project and the milestones achieved, gather feedback for future improvements and iterations

Thank You!

Questions?

Appendix

- [Project Proposal Link]
- [Project Proposal Slides]
- [Process Tracker]



Sample Tesseract Output

- **Step 1: Conduct a Manual Inspection of Attributes in the Selected PDFs**
  - The MLDS team will meticulously review the final 30 PDF documents to identify the most accurate ground truth values for each attribute. This will serve as our reference benchmark for subsequent validation checks
- **Step 2: Compare the Model Output with Ground Truth**
  - In-depth grading rubric will be developed to determine partial accuracy scores
- **Step 3: Quantify Model Accuracy Across Documents**
  - To determine the model's accuracy, divide the number of attributes the model correctly identified by the total number of attributes across all documents

Success Framework

- **Step 1: Conduct a Manual Inspection of Attributes in the Selected PDFs**
  - The MLDS team will meticulously review the final 30 PDF documents to identify the most accurate ground truth values for each attribute. This will serve as our reference benchmark for subsequent validation checks
- **Step 2: Compare the Model Output with Ground Truth**
  - In-depth grading rubric will be developed to determine partial accuracy scores
- **Step 3: Quantify Model Accuracy Across Documents**
  - To determine the model's accuracy, divide the number of attributes the model correctly identified by the total number of attributes across all documents

Final Deliverables

- **Deck on Document Extraction: Technology Comparison & Framework**
  - Pros and cons comparison of existing technologies
  - Suggested document extraction framework
  - Potential next steps
- **LLM Prompt Engineering**
  - Explore best system message, questions format, and LLM parameters (in progress)
- **WebApp Link & Demo**
  - Interactive conversational UI with the ability to chat with an uploaded documents
  - Run document intelligence AI across a collection of documents and store extracted features in structured format.

Potential Blockers

- **Tesseract and GPT 4.0 approvals**
  - Internal Cisco review process may dictate which solutions are viable
- **LLM Costs**
  - Costs could balloon with the addition of document chat functionality, by-attribute querying, and size of the final dataset (ex. applying the pipeline to thousands of documents)

GPT API Costs

Model	Input	Output
GPT 3.5 Turbo - 4k context	$0.0015 per 1K tokens	$0.002 per 1K tokens
GPT 3.5 Turbo - 16k context	$0.003 per 1K tokens	$0.004 per 1K tokens
GPT 4.0 - 8K context	$0.03 per 1K tokens	$0.06 per 1K tokens
GPT 4.0 - 32K context	$0.06 per 1K tokens	$0.12 per 1K tokens

Proposed Methods (From Midpoint)

Based on our testing , here is our proposed methodology:

- **Attribute Extraction Pipeline:** extracts a set list of attributes along with their context/proof in JSON format. Attributes are then added to an existing database or filestore system
  - Solutions: GPT 3.5, GPT 4.0 with or without RAG, running on an EC2 or ECS instance
- **Storage**: stores product attributes and their context for future querying
  - Solutions: Relational database or object storage (ex. S3 bucket)
- **Web App**: allows end users to upload documents or directory and extract information from the files.
  - Solutions: Streamlit
- **Chatbot**: allows users to ask questions about content from existing documents or submit new documents for querying purposes.
  - Solutions: LLM + Streamlit

Method Comparison - Example

Extraction Method: Pdfminer
Model: GPT 3.5
Data: gs-12-1565.pdf

One prompt for all attributes:

Answer: Supplier Name: AICC (Assumed based on the document header)

Product Name: EXAMAX2™ and EXAMEZZ2™ Connector System

Product Type: Connector System

Dimensions: The document does not explicitly state the dimensions of the connector system.

Orientation: The document mentions various orientations for different configurations, such as vertical header (VH), right angle receptacle (RAR), right angle orthogonal header (RAOH), and right angle header (RAH). Specific orientations may vary depending on the configuration.

Current Rating: The document refers to AICC Power Design Guide GS-20-0456 for the current rating. This guide should provide the relevant information.

Voltage: The document specifies the operating voltage rating as follows:
- Agency Voltage Rating < 30 VAC RMS / DC
- Non-Agency Voltage Rating = 150 VAC RMS or VDC Maximum

Frequency: The document does not mention a specific frequency.

Impedance / Capacitance: The document does not specify impedance or capacitance values.

Temperatures: The document provides the following temperature ranges:
- Operating Temperature Range = -55 to 85 degrees Celsius
- High Temperature Life Test Temperature = 85 degrees Celsius
- Thermal Shock Test Temperature Range = -55 to +85 degrees Celsius
- Humidity Test Temperature Range = 25 to 65 degrees Celsius

Please note that for specific technical details and performance requirements, it is recommended to consult the AICC Power Design Guide GS-20-0456 or other relevant documentation provided by the supplier.

Method Comparison - Mosaic Example

Extraction Method: PdfPlumber
Model: Mosaic-mpt7b-instruct
Data: gs-12-1565.pdf
KGM15CR51E106K-DATA.pdf

Document Name	KGM15CR51E106K-DATA.pdf	gs-12-1565.pdf
Document Length	1 page long	12 page long
Model Output	The electrical characteristics data for this capacitor are as follows: **Supplier name:** KEMET **Part number:** CM105X5R106K25A **Product type:** Ceramic Capacitors **Case size:** 0603 **Lead spacing:** X5R **Dimensions:** 1608(JIS)/0603(EIA)	Chained Grit Channels Chose Extensive EXCLUSIVE EX Chasers Chancing Chract Chrut Chrst Exact Chasing Exposed Ch Chrit Ch Ch Ch Ch Expose Ch Ch Exciting Ch Ex Ch Ch Ch Ch Ch Ch Exposing Ch Ch Ex Ch Ch Ch Ex Ch Ch Ch Ch
Response Time	3 seconds	3 minutes

Prompt Engineering [other prompts we tested]

Sample Prompts	Observations
"I'm looking for information in this document. Please extract the following information: You can **either provide the extracted information or provide the range of values** for the following: (a) Supplier Name, (b) Product type, (c) Dimensions, (d) Orientation if any Current Rating, (e) Voltage, (f) Frequency, (g) Impedance, (h) Capacitance, (i) Temperatures, (j) Other configuration settings (all the information that is not covered in the above categories)"	● Output **texts** with a list of attributes and results. ● Able to find and provide some additional attributes information with a range of value (if provided in pdf) compared with the previous one.
System message: Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer. Prompt: Please **extract any sentence** about the following information and include note for each attribute you extracted. (a) supplier name,..., (j) Other configuration settings. Please provide the answer **in the following format:** { "attribute": "attribute Name", "value": "attribute value", "note": "context to where and why you extracted the value" }, { "attribute": "attribute Name", "value": "attribute value", "context": "context to where and why you extracted the value" }, etc. for Supplier Name, Product type, Dimensions and so on.	● Output **JSON** with dictionaries of attribute names, attribute values and contexts. ● GPT 4 is able to output in the instructed format, and provide context for where its answers come from (either section number or section title) ● e.g.{ "attribute": "Frequency", "value": "10 to 500 to 10 hertz", "note": "This value was extracted from the section 8.6 Vibration Sinusoidal – EIA 364 -28 where it specifies the Frequency Range for the vibration test." }

Method Comparison - LLM

Models	Pros	Cons	Decision
GPT3.5	● More accurate than open-source models	● Less accurate and advanced than GPT-4.0	We use as final model for this project considering the balance between cost and performance
GPT4.0	● More accurate than GPT 3.5 on complex tasks	● More expensive than GPT 3.5	Recommend to use for pdf reading if budget allows
GPT4.0-turbo	● Can directly read image through API	● New model and high cost	Recommend to use if needs to deal with highly unstructured image information
Mistral-7B	● Completely free and secure	● Unable to detect some information of the attributes	Not suggested
Mosaic-mpt7b (instruct)	● Completely free and secure and accepts very long token inputs	● Hard to convince the model to follow the exact instruction ● Bad performance	Not suggested