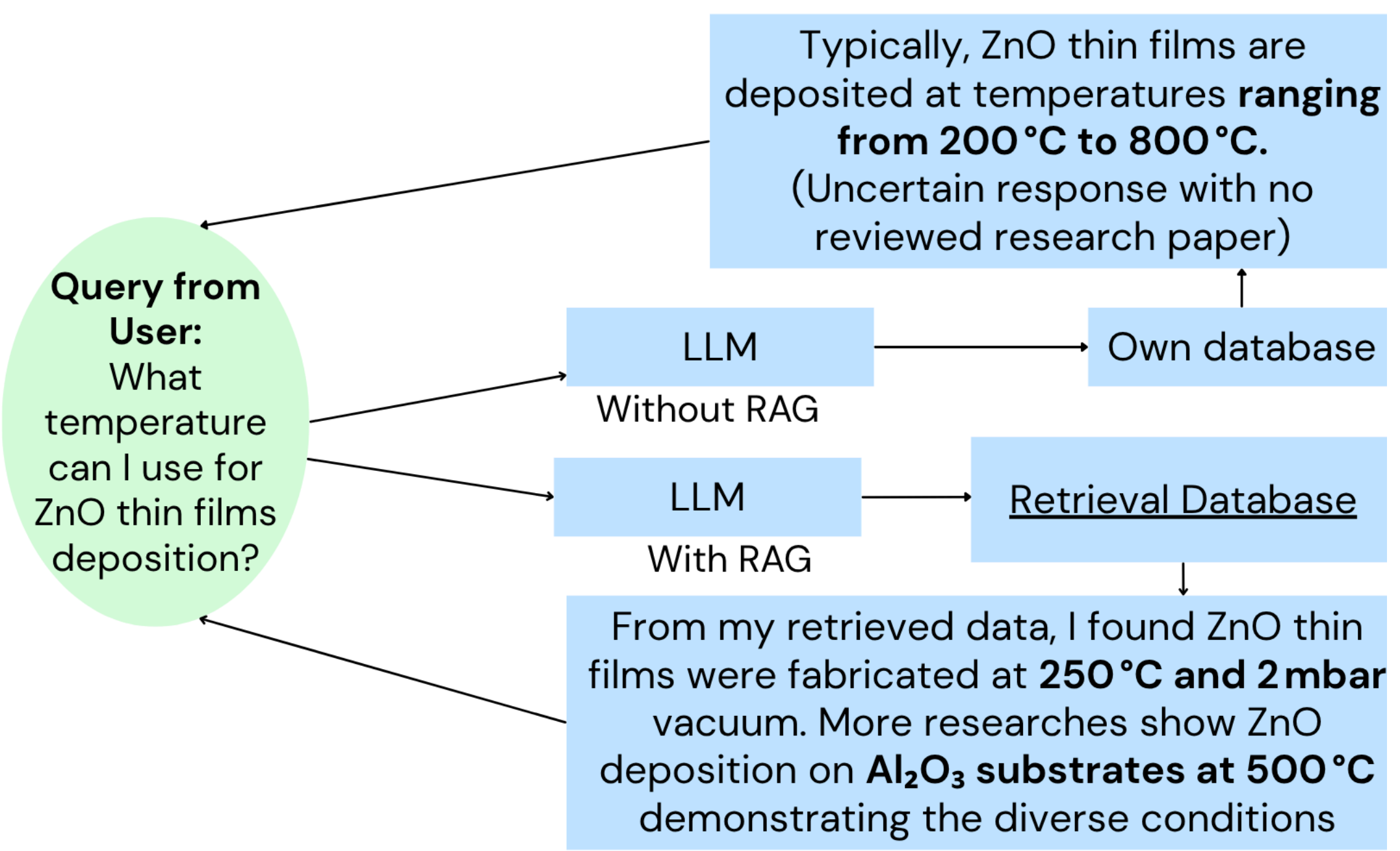
Automated Extraction of Synthesis Data from Unstructured Scientific Literature

Arsh Jaswal
Department of Chemical Engineering, IIT Kanpur
Prof. Salman Ahmad Khan
Department of Chemical Engineering IIT Kanpur

RAG in Scientific Literature

- Pulsed Laser Deposition, a highly precise and versatile thin film technique for film fabrication or material deposition.
- Manual scientific literature review and extraction for research parameters is an inefficient and slow process.
- Accurate retrieval is crucial to prevent hallucinated or misleading output, especially in research and experiments.
- NLP with RAG enhances experimental data retrieval.

How RAG works?


Query from User: What temperature can I use for ZnO thin films deposition?

LLM
Without RAG
Own database

Typically, ZnO thin films are deposited at temperatures **ranging from 200 °C to 800 °C.** (Uncertain response with no reviewed research paper)

LLM
With RAG
Retrieval Database

From my retrieved data, I found ZnO thin films were fabricated at **250 °C and 2 mbar** vacuum. More researches show ZnO deposition on **Al_2O_3 substrates at 500 °C** demonstrating the diverse conditions

Tools Used

sentence-transformers
pinecone-client
selenium
beautifulsoup4
undetected-chromedriver
requests

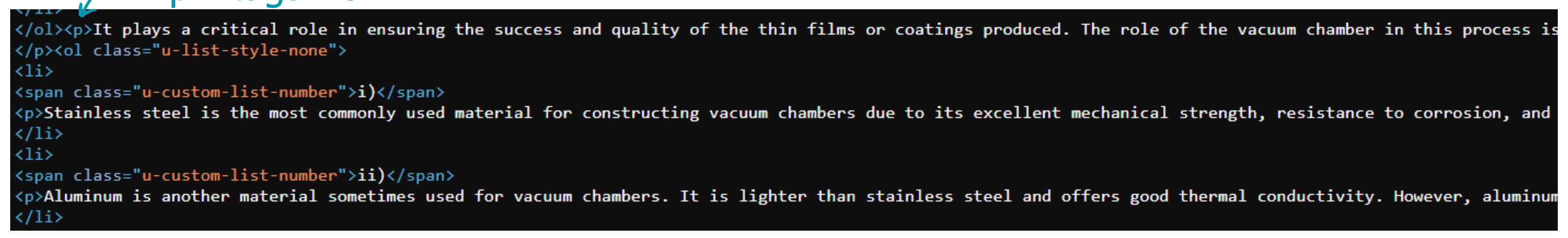
Google-generativeai
Gemini 1.5 flash API
Pincone DB API
Springer API
ModernBERT
MatSciBERT

numpy==1.24.4
matplotlib
scikit-learn
Ollama
LLaMA
Phi-3 mini

Paragraph Extraction

- Extracted <p> tags from HTML full texts
- Filtered out non-essential content (e.g., headers, references)
- Generalized for every topic; Generalized HTML parser

<p> tags from HTML



```
</ol><p>It plays a critical role in ensuring the success and quality of the thin films or coatings produced. The role of the vacuum chamber in this process is
</p><ol class="u-list-style-none">
<li>
<span class="u-custom-list-number">i)</span>
<p>Stainless steel is the most commonly used material for constructing vacuum chambers due to its excellent mechanical strength, resistance to corrosion, and
</li>
<li>
<span class="u-custom-list-number">ii)</span>
<p>Aluminum is another material sometimes used for vacuum chambers. It is lighter than stainless steel and offers good thermal conductivity. However, aluminum
</li>
```

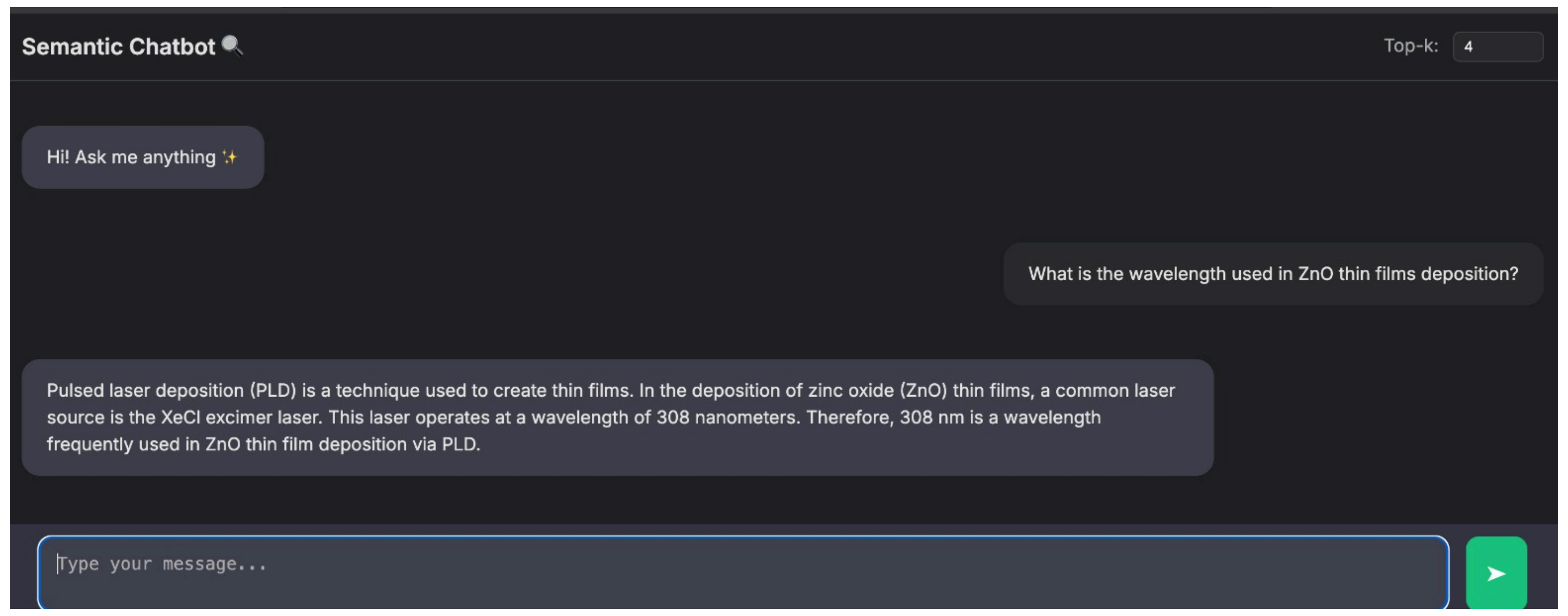
Sources:

1. Springer API (https://dev.springernature.com)
2. Automated springer webscraper (Please stick to website's user policy)

Pulsed Laser Deposition (a case study):

1. Scraped and Parsed about 4000 paragraphs related to Pulsed Laser Deposition (PLD)
2. Classifier using TF-IDF vectorizer and Logistic Regression classifier model configuration
3. Augmented with external 1000+ paragraphs, all marked 1
4. Choose top-k paragraph according to the need and the depth of the chatbot response

Chatbot Interface



Hi! Ask me anything ✨

What is the wavelength used in ZnO thin films deposition?

Pulsed laser deposition (PLD) is a technique used to create thin films. In the deposition of zinc oxide (ZnO) thin films, a common laser source is the XeCl excimer laser. This laser operates at a wavelength of 308 nanometers. Therefore, 308 nm is a wavelength frequently used in ZnO thin film deposition via PLD.

Retrieval System

Embedder used: ModernBERT, MatSciBERT; ModernBERT outperformed MatSciBERT
Database of 9000+ paragraphs, all related to PLD experiements and related researches
Cloud based vector database (pinecone.io) storing ModernBERT embeddings

Semantic Search:

1. Query Expansion
2. Single Semantic Search ranking

Deep Search:

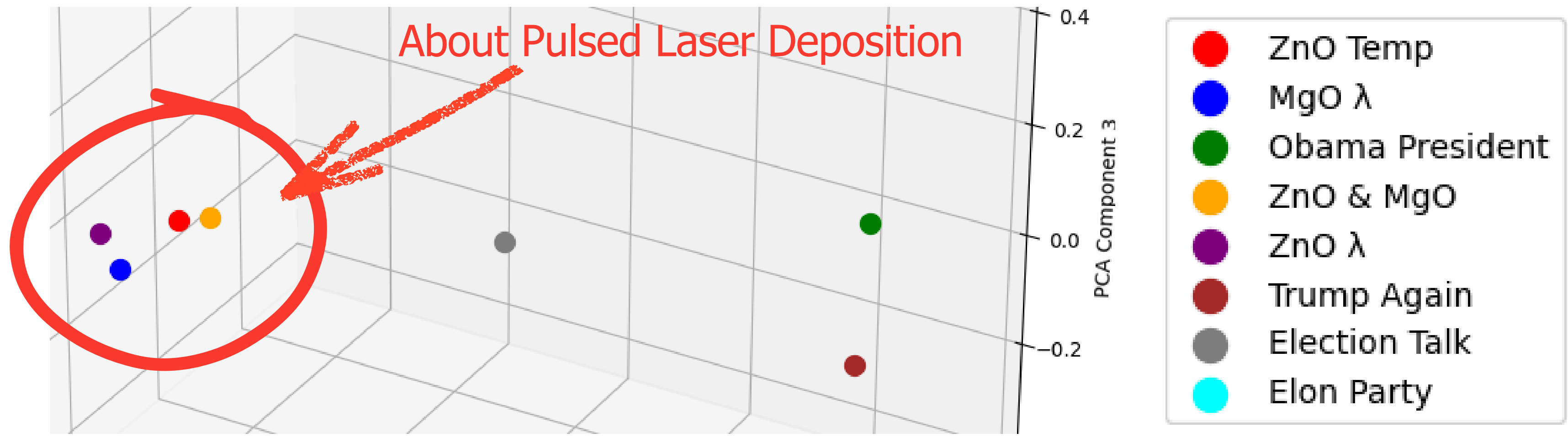
Retrieval Precision Enhancement (with history):

- Query Expansion/Rewriting
- Filter pruning (kw_score(p_i))
- Heuristic Reranking (kw_score(p_i) > kw_score(p_{i+1}))
- Multi-Semantic Reranking

Filter Pruning

$$\mathrm{kw\_score}(p_i) = \sum_{k \in K(x')} \mathbf{1}_{k \in p_i}$$

$$\mathbf{1}_{k \in p_i} = \begin{cases} 1 & \text{if keyword } k \text{ is in } p_i \\ 0 & \text{otherwise} \end{cases}$$

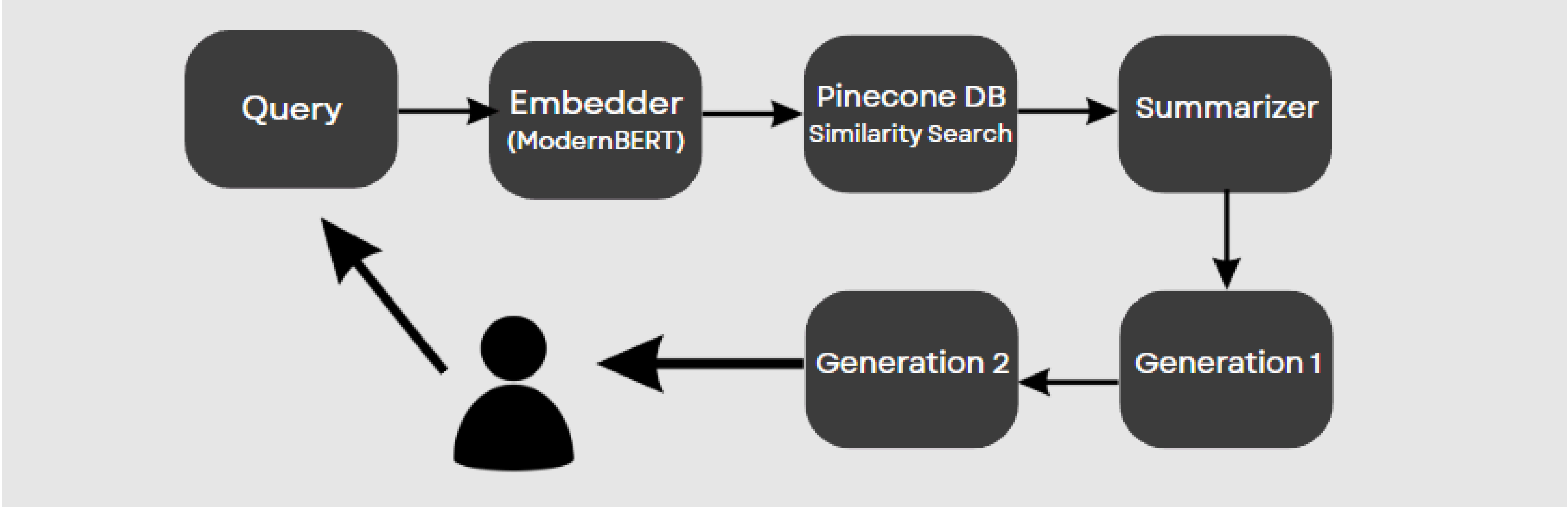


3D representation of n dimension ModernBERT embedding after dimensional reduction

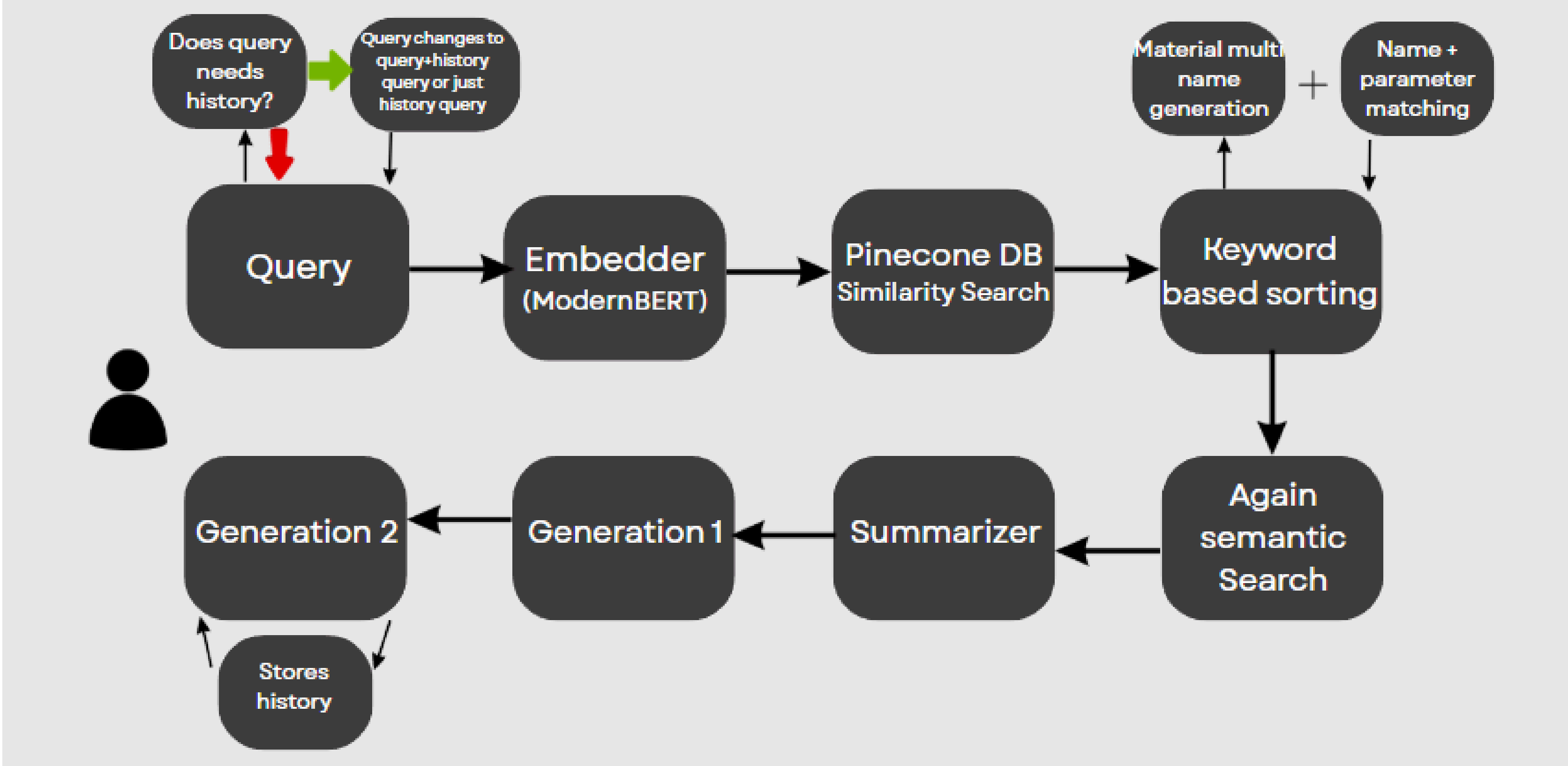
Generation Pipelines

Utilizes Retrieval Augmented Generation
Leverages Gemini-1.5 flash with Pinecone DB

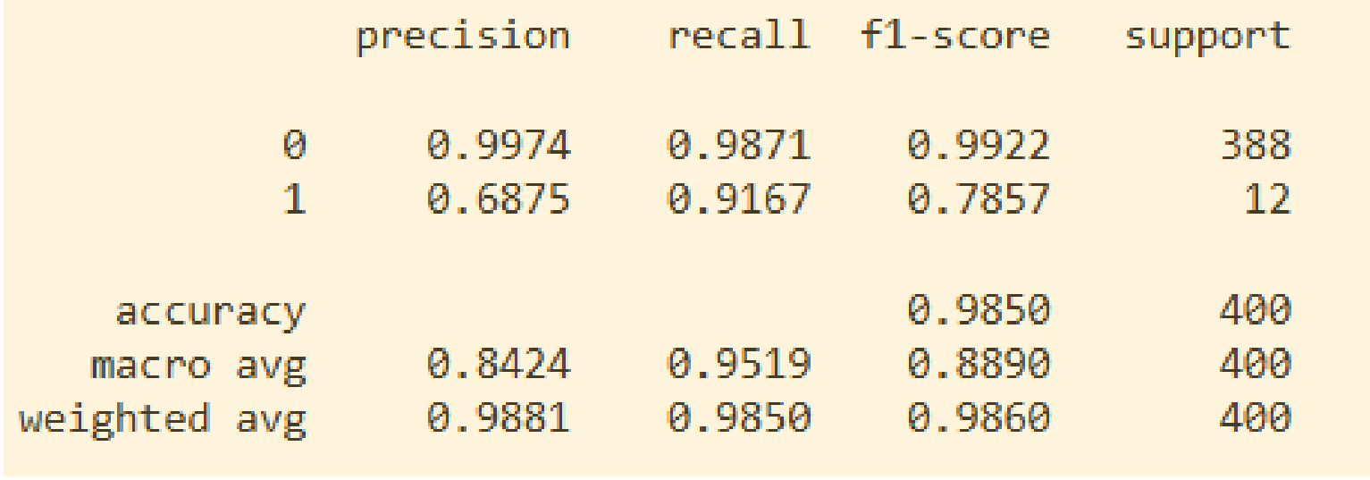
Semantic Only:



Heuristic Reranking + Filter Pruning + Query-Expansion + Multi-semantic reranking Deep Search:


Does query needs history?
Query changes to query+history query or just history query
Query
Embedder (ModernBERT)
Pinecone DB Similarity Search
Keyword based sorting
Material multi name generation
+
Name + parameter matching
Again semantic Search
Summarizer
Generation 1
Generation 2
Stores history

Results



```
              precision    recall  f1-score   support

           0     0.9974    0.9871    0.9922       388
           1     0.6875    0.9167    0.7857        12

    accuracy                         0.9850       400
   macro avg     0.8424    0.9519    0.8890       400
weighted avg     0.9881    0.9850    0.9860       400
```

Classifier before augmentation

Classifier after augmentation

Class	Precision	Recall	F1-Score	Support
0.0	0.9764	0.9764	0.9764	423
1.0	0.9558	0.9558	0.9558	226
Accuracy			**0.9692**	649
Macro Avg	0.9661	0.9661	0.9661	649
Weighted Avg	0.9692	0.9692	0.9692	649

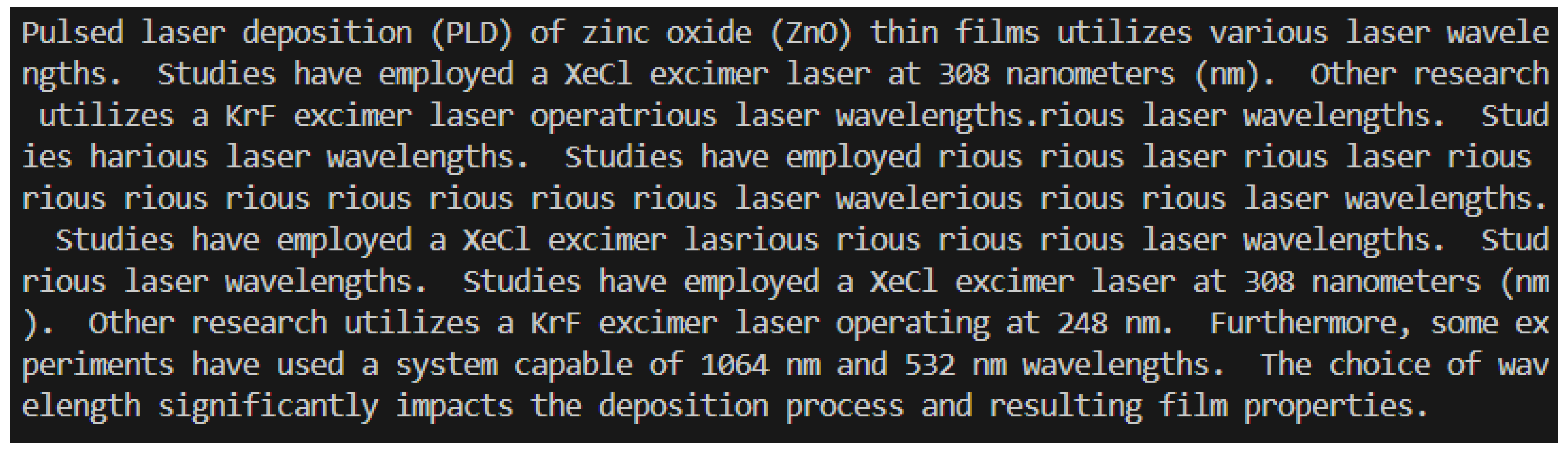
AUC-ROC: 0.99742268041

	Predicted 0	Predicted 1
Actual 0	413	10
Actual 1	10	216

Confusion Matrix after augmentation

Semantic Search:

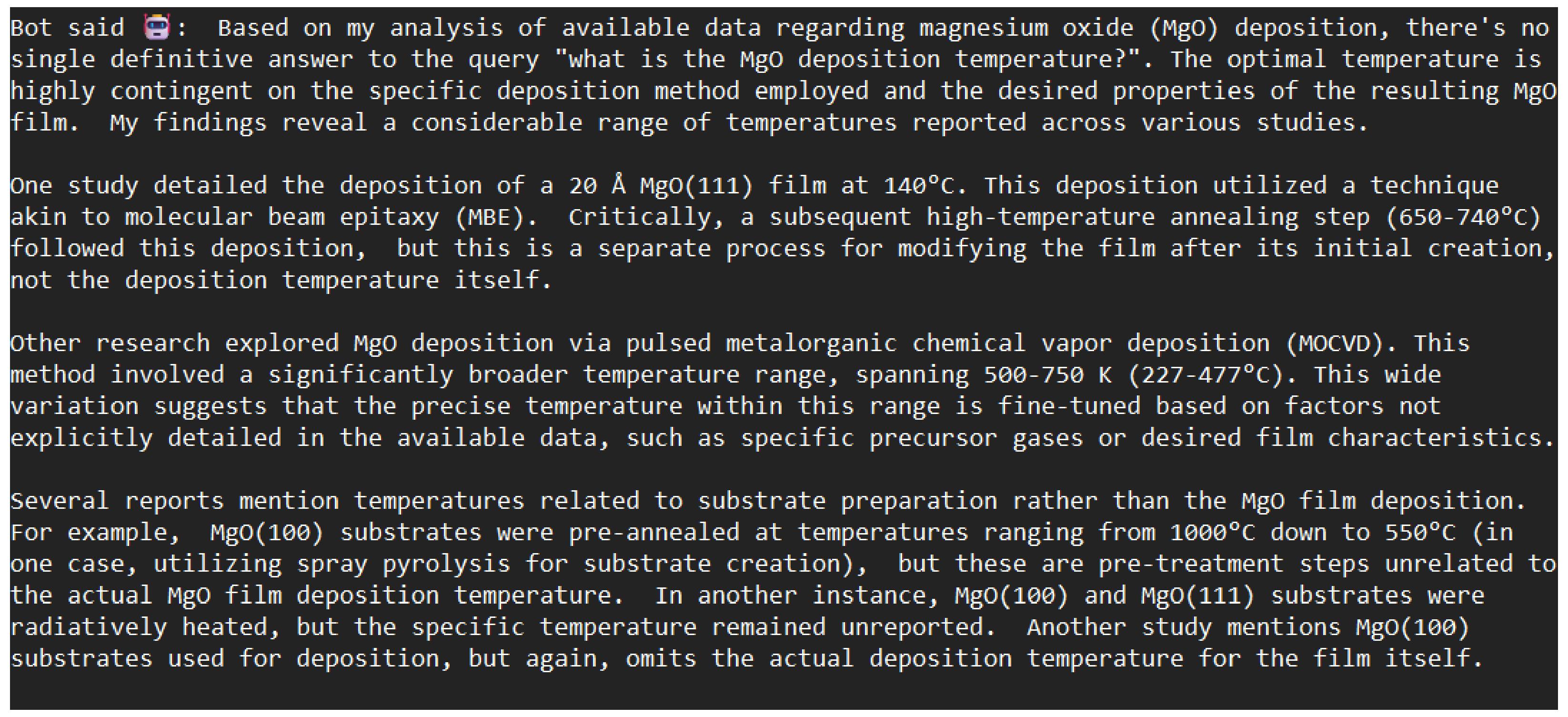
Example Query: What is the wavelength used in ZnO thin films?



```
Pulsed laser deposition (PLD) of zinc oxide (ZnO) thin films utilizes various laser wavele
ngths.  Studies have employed a XeCl excimer laser at 308 nanometers (nm).  Other research
 utilizes a KrF excimer laser operatrious laser wavelengths.rious laser wavelengths.  Stud
ies harious laser wavelengths.  Studies have employed rious rious laser rious laser rious
rious rious rious rious rious rious rious laser wavelerious rious rious laser wavelengths.
  Studies have employed a XeCl excimer lasrious rious rious rious laser wavelengths.  Stud
rious laser wavelengths.  Studies have employed a XeCl excimer laser at 308 nanometers (nm
).  Other research utilizes a KrF excimer laser operating at 248 nm.  Furthermore, some ex
periments have used a system capable of 1064 nm and 532 nm wavelengths.  The choice of wav
elength significantly impacts the deposition process and resulting film properties.
```

Deep Search (Keyword enhanced Multi-Semantic):

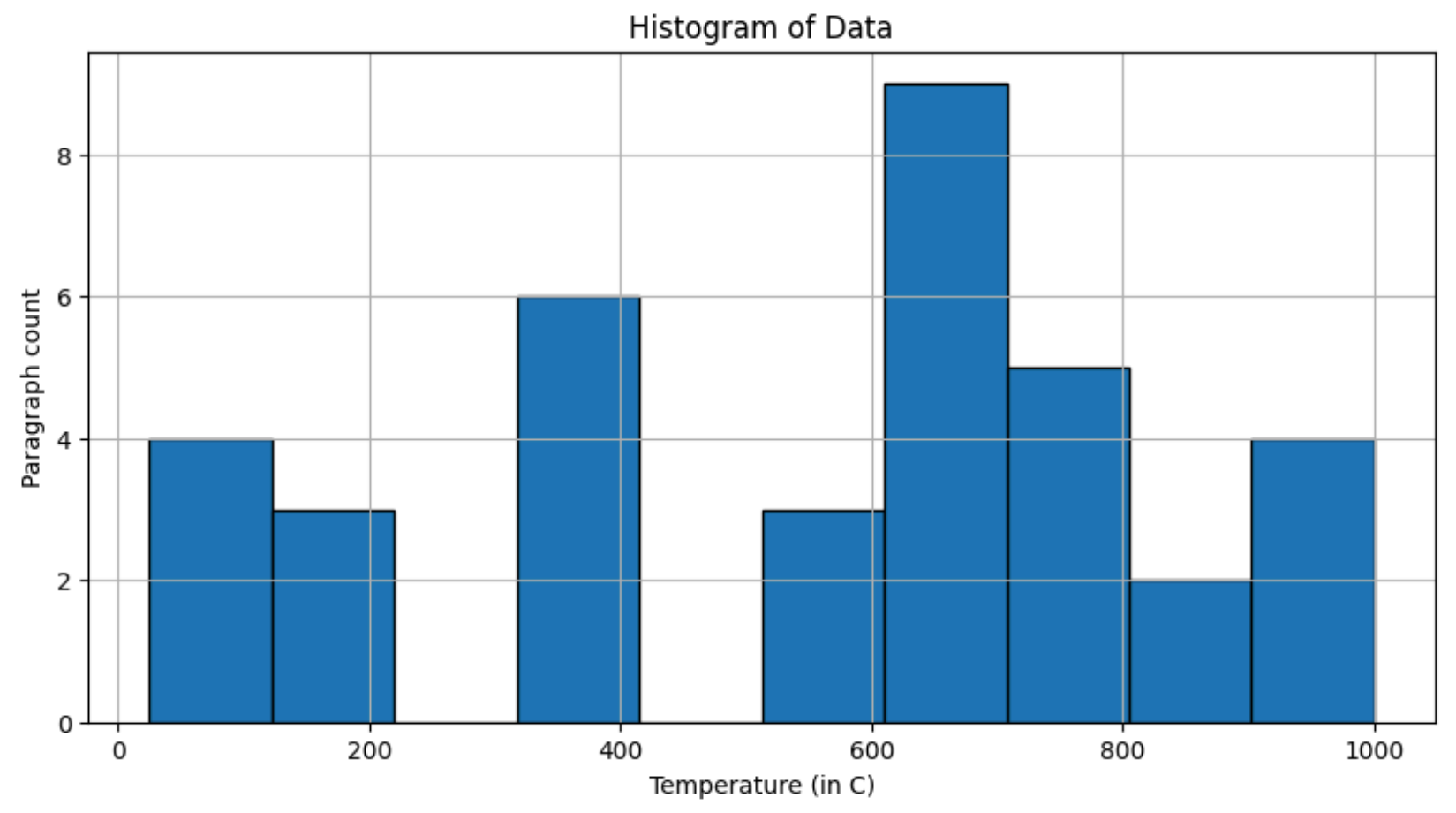
Example Query: What is the MgO deposition temperature?



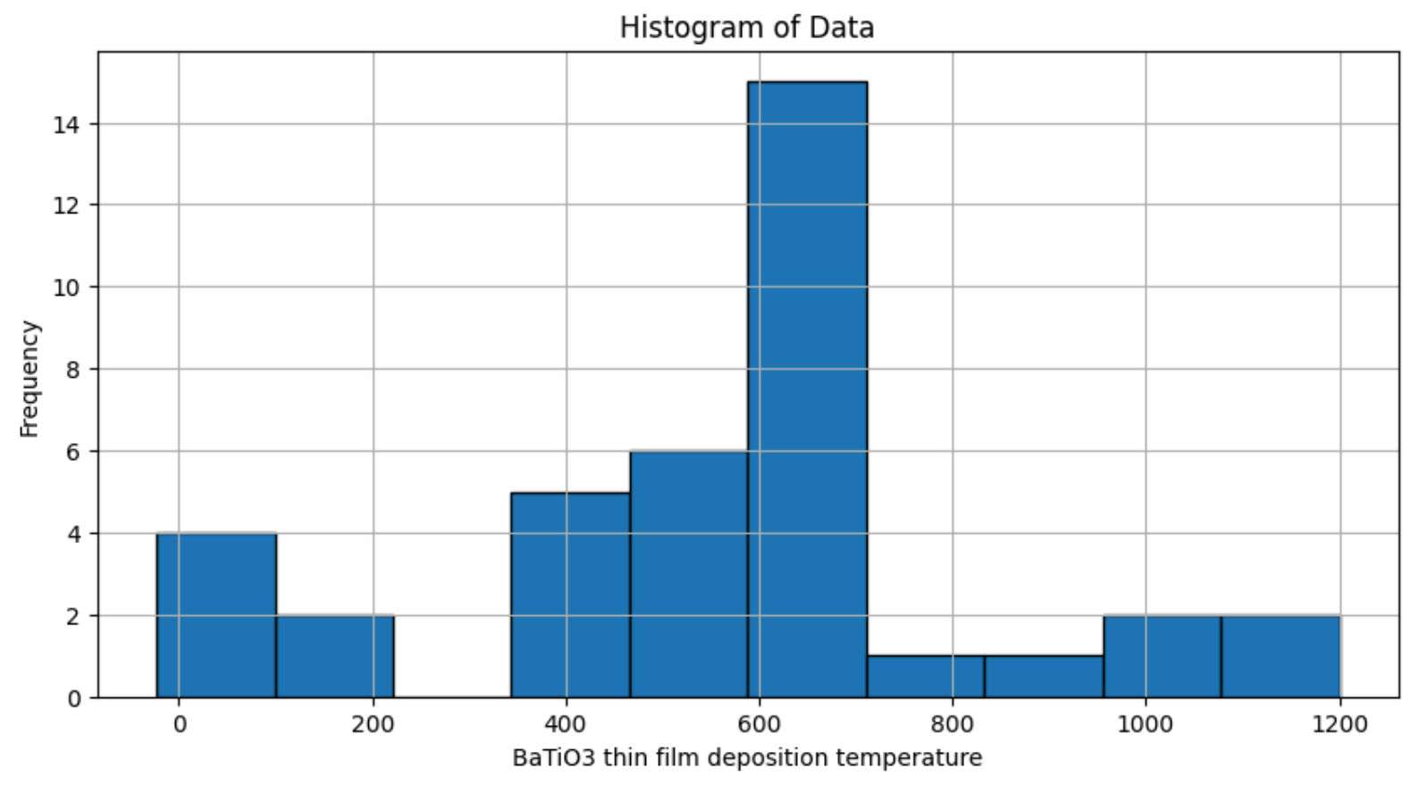
```
Bot said 🤖:  Based on my analysis of available data regarding magnesium oxide (MgO) deposition, there's no
single definitive answer to the query "what is the MgO deposition temperature?". The optimal temperature is
highly contingent on the specific deposition method employed and the desired properties of the resulting MgO
film.  My findings reveal a considerable range of temperatures reported across various studies.

One study detailed the deposition of a 20 Å MgO(111) film at 140°C. This deposition utilized a technique
akin to molecular beam epitaxy (MBE).  Critically, a subsequent high-temperature annealing step (650-740°C)
followed this deposition,  but this is a separate process for modifying the film after its initial creation,
not the deposition temperature itself.

Other research explored MgO deposition via pulsed metalorganic chemical vapor deposition (MOCVD). This
method involved a significantly broader temperature range, spanning 500-750 K (227-477°C). This wide
variation suggests that the precise temperature within this range is fine-tuned based on factors not
explicitly detailed in the available data, such as specific precursor gases or desired film characteristics.

Several reports mention temperatures related to substrate preparation rather than the MgO film deposition.
For example,  MgO(100) substrates were pre-annealed at temperatures ranging from 1000°C down to 550°C (in
one case, utilizing spray pyrolysis for substrate creation),  but these are pre-treatment steps unrelated to
the actual MgO film deposition temperature.  In another instance, MgO(100) and MgO(111) substrates were
radiatively heated, but the specific temperature remained unreported.  Another study mentions MgO(100)
substrates used for deposition, but again, omits the actual deposition temperature for the film itself.
```

Histogram Plotter:

Query: MgO thin films temperature



Query: BaTiO3 thin films temperature



References

- https://arxiv.org/html/2410.19572
- https://doi.org/10.1038/s41467-024-45563-x
- https://doi.org/10.1039/D4DD00051J

Students-Undergraduate Research Graduate Excellence

Centre for Continuing Education, IIT Kanpur, India

Contact : arshj23@iitk.ac.in