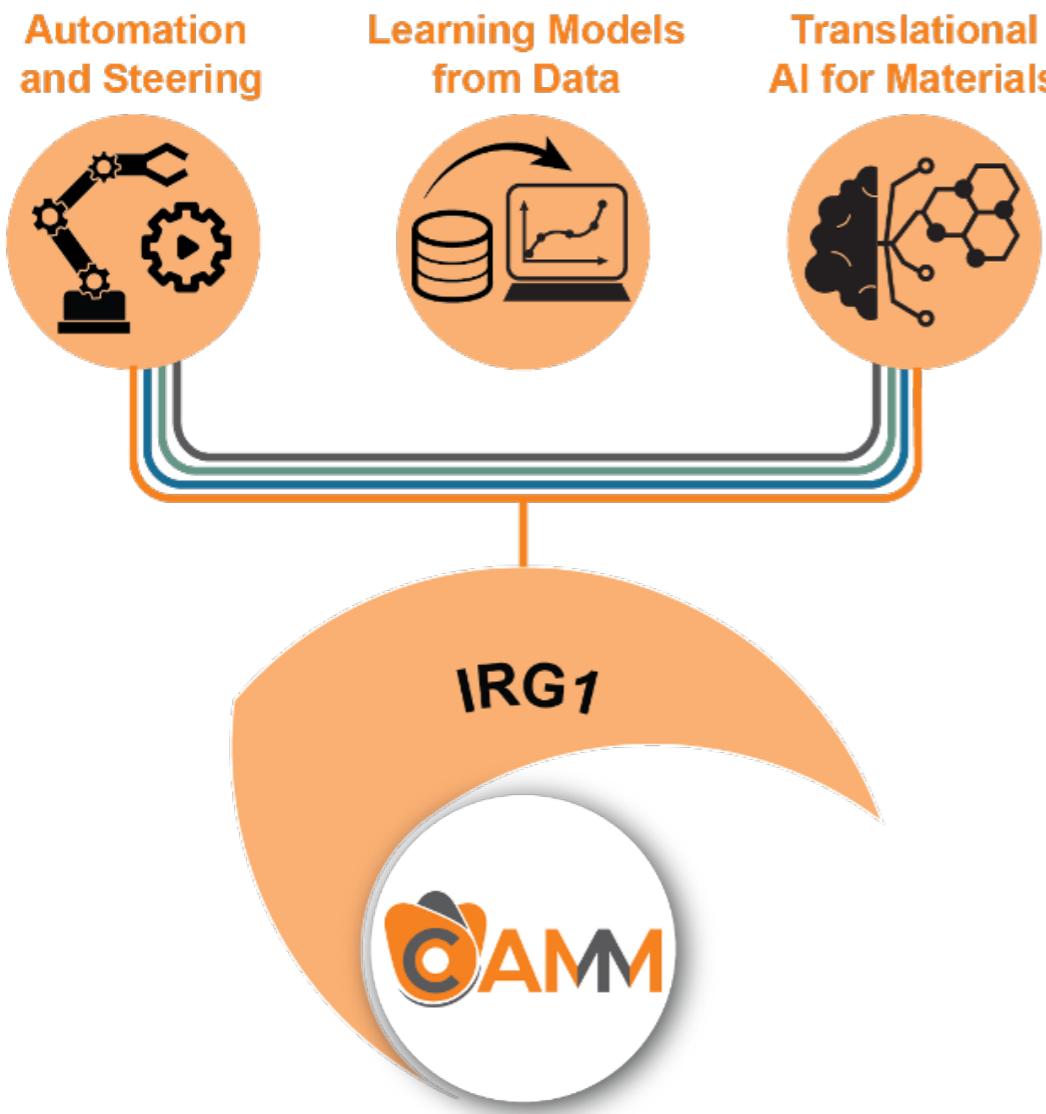


# Large language models for interpretable materials databases

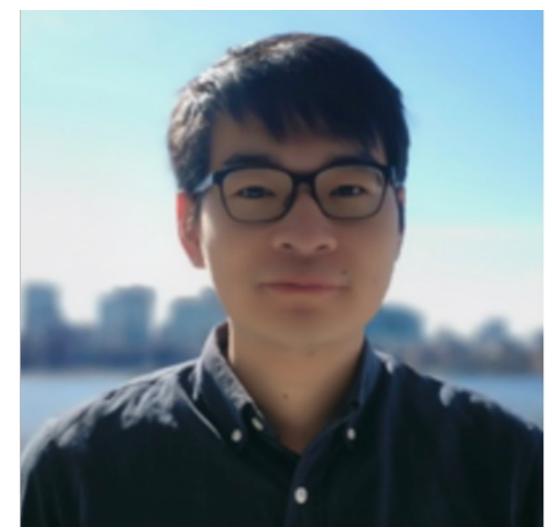
Louis Primeau  
May 3rd, 2024



# Collaborations



Adrian Del Maestro

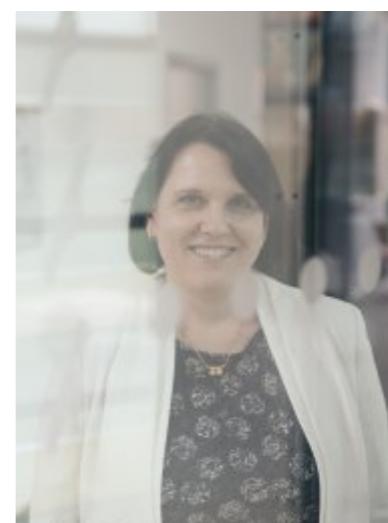


Yang Zhang

Experimental groups



David Mandrus



Claudia Felser



Haidong Zhou

Goal 1: Create **machine**-readable materials databases  
(structured data) about materials.

- Extract standard material properties.
- Extract synthesis / fabrication methods.
- Perform machine learning on this database

Goal 2: Create **human**-readable literature report (LLM wiki)

- Provide all relevant measured properties (referenced) to a given material / material class.
- Summarize above properties into a human-readable format.

Goal 1: Identify material-parameter and synthesis method databases for materials.

Problem: Scientific papers are a very bad way of reporting results - they constitute unstructured and diverse datasets

Sub-problems:

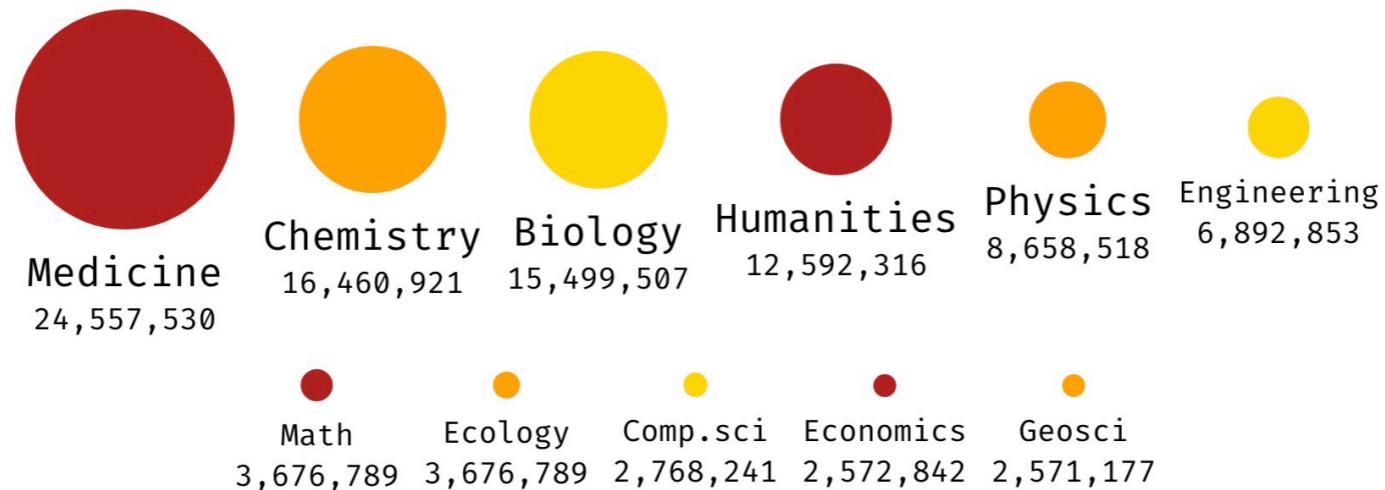
- Which papers should we download?
- Extraction of text and figures from old scientific papers for which no latex is available
- Extraction of information from text, table and figures

# What do we have?

- Hand-labeled database: MDR SuperCon, ICSD, CCDC....
- 4x A6000 GPUs (48GB memory/card), for 8B Llama 3

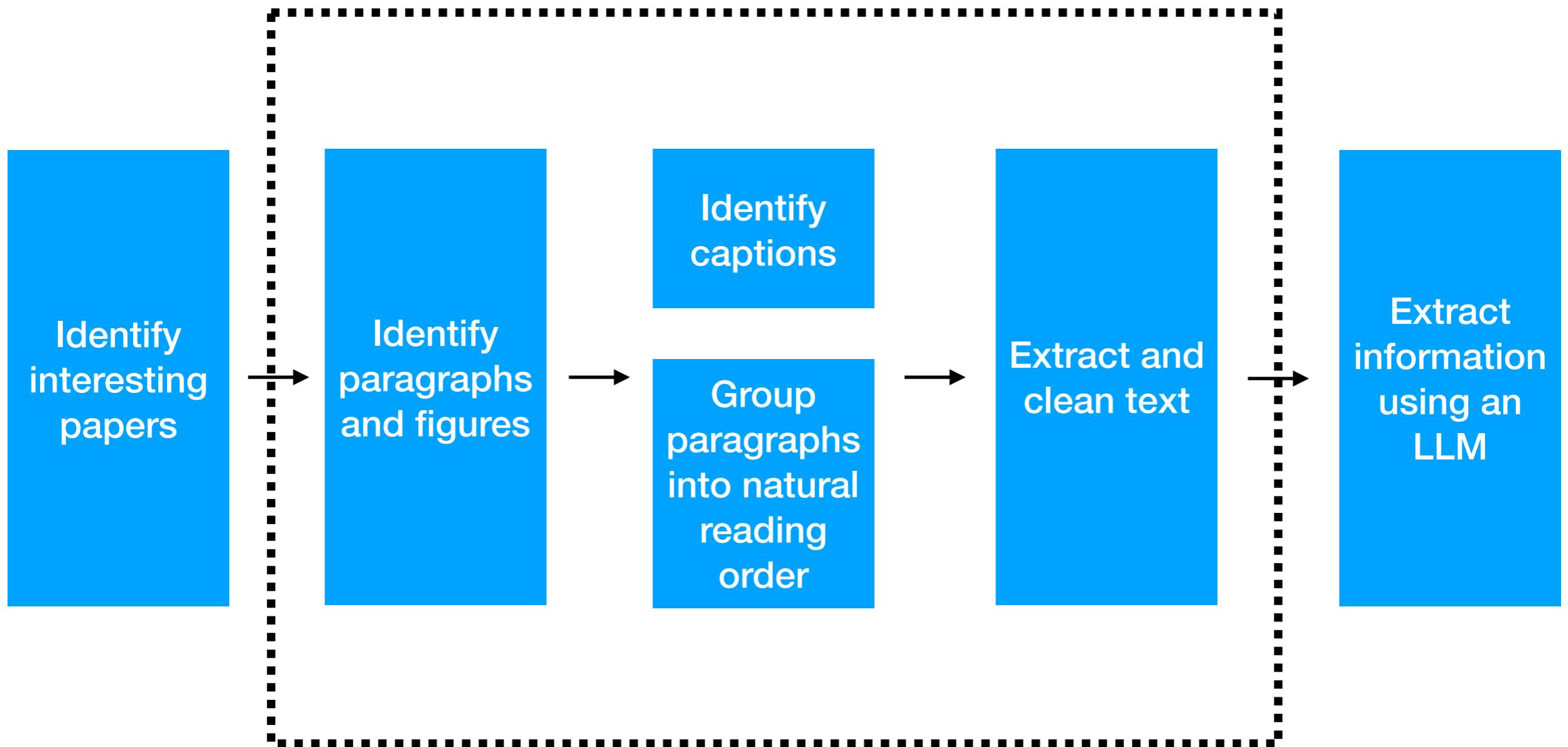
32x H100 GPUs (80 GB /card) for 70B Llama 3, 768x A100 for tuning

- Research Papers (~100 TB for major scientific publishers)



- Downloaded: arXiv (~ 1.1 TB for 2.4m papers), 1400 from 3 groups, PRB for SC (~2000 papers, well structured) + 2000 magnetic material papers

# Document Understanding



# Pipeline Results

PHYSICAL REVIEW B 71, 134526 (2005)

## Dimensionality of superconductivity and vortex dynamics in the infinite-layer cuprate $\text{Sr}_{0.9}\text{M}_{0.1}\text{CuO}_2$ ( $\text{M}=\text{La},\text{Gd}$ )

V. S. Zapf,<sup>1,2</sup> N.-C. Yeh,<sup>1</sup> A. D. Beyer,<sup>1</sup> C. R. Hughes,<sup>1</sup> C. H. Mielke,<sup>2</sup> N. Harrison,<sup>2</sup> M. S. Park,<sup>3</sup> K. H. Kim,<sup>3</sup> and S.-I. Lee<sup>3</sup>

<sup>1</sup>Department of Physics, California Institute of Technology, Pasadena, California, USA

<sup>2</sup>National High Magnetic Field Laboratory, Los Alamos, New Mexico 87545, USA

<sup>3</sup>Department of Physics, Pohang University of Science and Technology, Pohang, Korea

(Received 21 April 2004; revised manuscript received 9 December 2004; published 29 April 2005)

The high magnetic-field phase diagram of the electron-doped infinite layer high-temperature superconducting (high- $T_c$ ) compound  $\text{Sr}_{0.9}\text{La}_{0.1}\text{CuO}_2$  was probed by means of penetration depth and magnetization measurements in pulsed fields to 60 T. An anisotropy ratio of 8 was detected for the upper critical fields with  $H$  parallel ( $H_{c2}^{ab}$ ) and perpendicular ( $H_{c2}^c$ ) to the  $\text{CuO}_2$  planes, with  $H_{c2}^{ab}$  extrapolating to near the Pauli paramagnetic limit of 160 T. The longer superconducting coherence length than the lattice constant along the  $c$  axis indicates that the orbital degrees of freedom of the pairing wave function are three dimensional. By contrast, low-field magnetization and specific heat measurements of  $\text{Sr}_{0.9}\text{Gd}_{0.1}\text{CuO}_2$  indicate a coexistence of bulk  $s$ -wave superconductivity with large moment Gd paramagnetism close to the  $\text{CuO}_2$  planes, suggesting a strong confinement of the spin degrees of freedom of the Cooper pair to the  $\text{CuO}_2$  planes. The region of the magnetic field-temperature phase diagram between  $H_{c2}^{ab}$  and the irreversibility line in the magnetization,  $H_{\text{irr}}^{ab}$ , in  $\text{Sr}_{0.9}\text{La}_{0.1}\text{CuO}_2$  is anomalously large for an electron-doped high- $T_c$  cuprate. The large reversible region even approaching zero temperature rules out thermal depinning scenarios. The temperature dependence of  $H_{\text{irr}}^{ab}$  also differs fundamentally from those predicted for the quenched-disorder-induced vortex phase transitions for  $H \parallel c$  at low temperatures. Thus, our finding of a strongly suppressed  $H_{\text{irr}}^{ab}$  relative to  $H_{c2}^{ab}$  at low temperatures is suggestive of the existence of additional quantum fluctuations, possibly due to a magnetic-field-induced competing order such as the spin-density wave (SDW).

DOI: 10.1103/PhysRevB.71.134526

PACS number(s): 74.25.Dw, 74.25.Op, 74.72.Dn, 74.25.Ha

### I. INTRODUCTION

In the high- $T_c$  cuprate superconductors, it has been suggested that anisotropy plays an important role in the superconducting pairing mechanism and the elevated  $T_c$  in both experimental and theoretical work.<sup>1,2</sup> It is surprising therefore to find superconductivity (SC) with  $T_c=43$  K in the optimal electron-doped infinite-layer cuprates  $\text{Sr}_{0.9}\text{M}_{0.1}\text{CuO}_2$  ( $\text{M}=\text{La},\text{Gd}$ ), which exhibit only a 16% difference between the  $a$  and  $c$  tetragonal lattice parameters. The structure of  $\text{Sr}_{0.9}\text{M}_{0.1}\text{CuO}_2$  is the most basic among all high- $T_c$  cuprates, consisting entirely of  $\text{CuO}_2$  sheets separated by rare-earth (RE) ions with tetragonal lattice parameters  $c=3.41$  Å and  $a=3.95$  Å.<sup>3</sup> The recent success in producing high-quality polycrystalline samples of the infinite-layer cuprates with no observable impurity phases<sup>3</sup> has engendered a renewed interest in these compounds. X-ray near-edge absorption spectroscopy indicate electron doping,<sup>4</sup> and bulk SC has been verified by powdered magnetization ( $M$ ) measurements<sup>5</sup> and specific heat ( $C$ ) measurements (data presented later in this work). Several recent studies of these high-purity polycrystalline samples suggest three-dimensional (3D) superconductivity in  $\text{Sr}_{0.9}\text{La}_{0.1}\text{CuO}_2$ . Scanning tunnel spectroscopy (STS) measurements<sup>6</sup> indicate an unconventional but isotropic

length ( $\xi_c$ ) from a Hao-Clem analysis<sup>10</sup> of the reversible magnetization of grain-aligned polycrystal and found that  $\xi_c$  exceeds the spacing between the  $\text{CuO}_2$  planes, indicating 3D superconductivity. On the other hand, they also find significant anisotropy between magnetic fields  $H \leq 5$  T oriented parallel and perpendicular to the  $\text{CuO}_2$  planes, with an anisotropy ratio  $\gamma = \xi_c/\xi_{ab} = H_{c2}^{ab}/H_{c2}^c = 9.3$ , which is larger than  $\gamma = 5$  observed in  $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$  but much smaller than  $\gamma = 55$  observed in optimally doped  $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8-\delta}$ .<sup>5,11–13</sup> It is interesting to note that the only major crystallographic difference between the  $a$ - $b$  and the  $c$  directions in  $\text{Sr}_{0.9}\text{La}_{0.1}\text{CuO}_2$  is the presence of oxygen in the  $a$ - $b$  plane, which allows coupling of adjacent Cu spins and has been implicated as the cause of antiferromagnetic ordering or spin fluctuations in other members of the high- $T_c$  cuprate family, as well as a possible mechanism for superconducting pairing. The importance of the  $\text{CuO}_2$  planes to the SC in  $\text{Sr}_{0.9}\text{M}_{0.1}\text{CuO}_2$  is further supported by the fact that Ni substitution on the Cu site rapidly suppresses  $T_c$ , whereas out-of-plane Gd substitution on the Sr site leaves  $T_c$  unchanged.<sup>7,14</sup>

In this work we determine the upper critical field  $H_{c2}$  and the irreversibility field  $H_{\text{irr}}$  of  $\text{Sr}_{0.9}\text{La}_{0.1}\text{CuO}_2$  by means of magnetization and penetration depth measurements in pulsed magnetic fields up to 60 T in order to directly investigate the degrees of supercritical finite- $T$  anisotropy and the role of van

'V. S. Zapf, .2 N.-C. Yeh, A. D. Beyer, C. R. Hughes, C. H. Mielke,2 N. Harrison,2 M. S. Park, K. H. Kim,3 and S.-I. Lee\n z\n DOI: 10.1103/PhysRevB.71.134526 PACS number(s): 74.25.Dw, 74.25.0p, 74.72.Dn, 74.25.1a length (E) from a Hao-Clem analysis\'' of the reversible magnetization of grain-aligned polycrystal and found that fa exceeds the spacing between the Cu0, planes, indicating 3D superconductivity. On the other hand, they also find significant anisotropy between magnetic fields Hs5 T oriented parallel and perpendicular to the Cu02 planes, with an anisotropy ratio y= \$\backslash xi\_{c}/\backslash xi\_{ab}\backslash !=\backslash !H\_{c2}^{ab}/H\_{c2}^c = 9.3\$, which is larger than 2 = 5 observed in YBa,Cu.0\$7-\backslash delta\$ nasnsadnaosals \$\backslash mathrm{It}\$ is in observed in optimally doped BisSr,CaCu20s-8\nteresting to note that the only major crystallographic difference between the a-b and the c directions in Sro.,La, 1Cu0 is the presence of oxygen in the a-b plane, which allows coupling of adjacent Cu spins and has been implicated as the cause of antiferromagnetic ordering or spin fluctuations in

# What about ChatGPT-4o?



PhysRevB.71.134526-2.pdf  
PDF

What is the material studied?



The material studied in the provided paper is the infinite-layer cuprate  $\text{Sr}_{0.9}\text{M}_{0.1}\text{CuO}_2$  where M is either La (Lanthanum) or Gd (Gadolinium). The paper investigates the superconductivity and vortex dynamics in this high-temperature superconducting compound [1].



- Struggles on poorly scanned old papers, for much the same cases as the pipeline - bad reading subscripts.

Superconductivity distance blah blah decreases gradually as its temperature is lowered, even down to near absolute zero, abcdef fasdfasdf a b c d e f g superconductor a b c d e  $H_5C_3L_{20}Co_{1-x}Be_x$  even down to near absolute zero We introduce a method to predict arbitrary, non-parametric probability distributions over the abcd hey there blah blah rotation manifold. This is particularly useful for pose estimation of symmetric and nearly symmetric objects, since the

# Practical LLMs

The vocabulary of LLMs:

Tokens - The tokens are what the LLM processes. Instead of feeding strings to the LLMs, the input text is ‘tokenized’- it is mapped (invertibly) to a sequence of numbers. LLama3 uses a BPE tokenizer, and can take up to 8000 tokens (~8000 words).

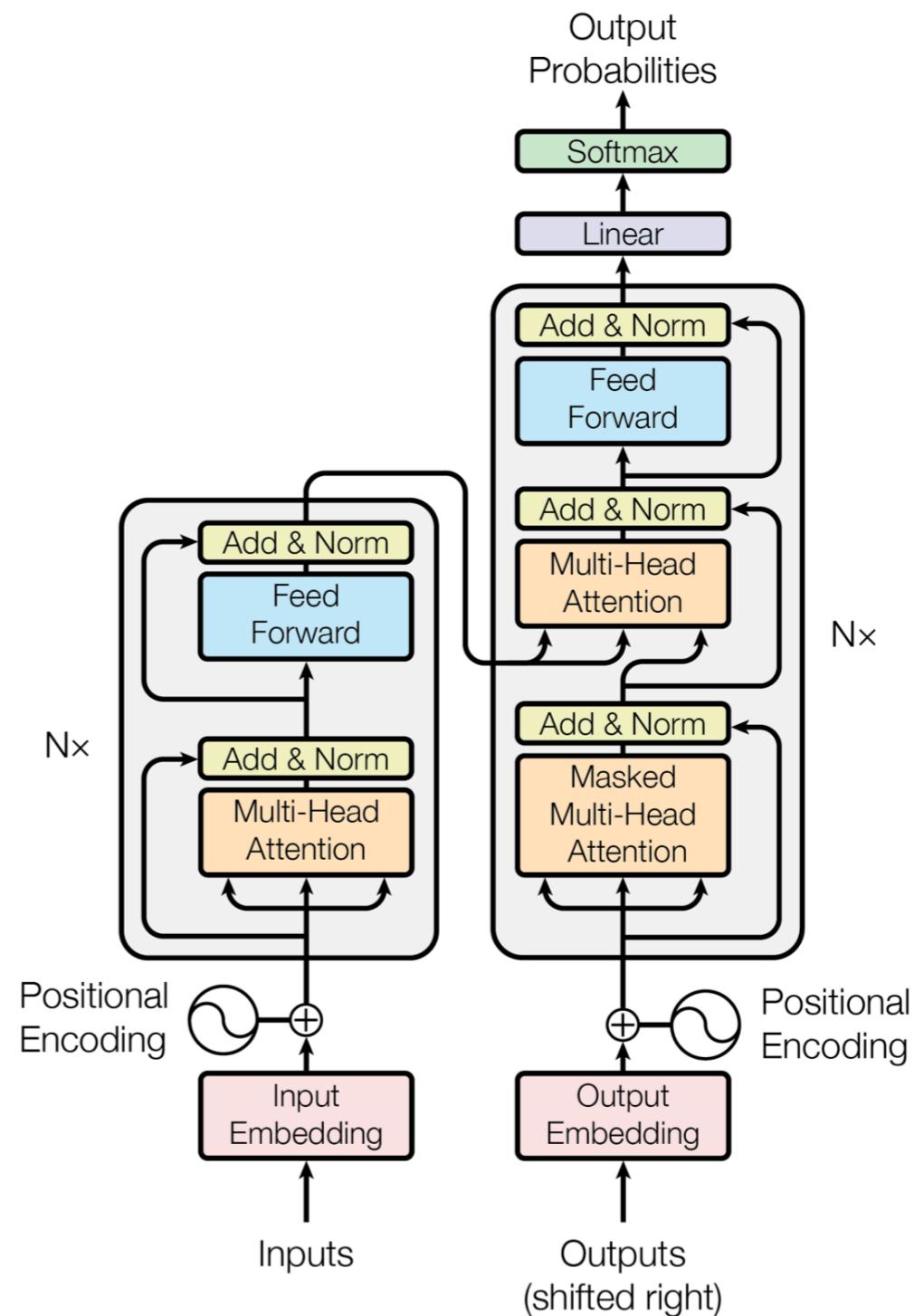
Embeddings - You can then map these embeddings into a latent space in which sentences with similar semantic content are close together. These are the embeddings. Here we can distinguish between two classes of models:

Sentence Transformers - Small and efficient at producing embeddings. These embeddings can be compared with cosine similarity.

LLMS - Can do the same thing above but are meant to generate responses. The embeddings are not necessarily meaningful.

# Transformers and Attention

- Compare the input query embeddings to the input keys which are the token embeddings in the vocabulary of the LLM using a learnable weight matrix.  
Repeat this a lot of times for learnability and long range correlations.
- Linear scaling in the sequence length, easy to parallelize on GPUs.



# GPTs: Llama 3

LLama3 is a transformer network trained on a corpus of 15T tokens.

Available in 8B and 70B parameter flavors. 8096 input token length (2x llama2 capabilities)



Uses various improvements to the transformer architecture since its introduction in 2017.

	Time (GPU hours)	Power Consumption (W)	Carbon Emitted(tCO2eq)
Llama 3 8B	1.3M	700	390
Llama 3 70B	6.4M	700	1900
Total	7.7M		2290

# Superconductivity Dataset

## MDR SuperCon Datasheet

Contains ~30000 human-labelled material entries, along with the paper they were published in.

We can use this as a training dataset. We obtain 1600 papers from physical review b.

### Problems:

- Each paper usually studies multiple compounds (from the same family).
- Not every paper lists the critical temperature, critical magnetic field, etc in the same way, or at all. In all, we extract 1200 papers that have entries listing some superconducting critical temperature.

# Prompts for Superconductivity Parameter Retrieval

What is the material studied in this paper? Format the answer as MATERIAL: {Chemical Formula}. If there are multiple materials, separate them with &. Just give a formula and do not provide any explanation. Here are some example outputs: 'MATERIAL: Ga & Al & TeS3 & UF6', 'MATERIAL: PrOs<sub>4</sub>Sb<sub>12</sub> & PrOs<sub>3</sub>Sb<sub>13</sub>'

What is the critical temperature at zero-field of {MATERIAL}? Just give a number and do not provide any explanation. Format the answer as CRITICAL TEMPERATURE: {Number} K. Here are some example outputs: 'CRITICAL TEMPERATURE: 3K', 'CRITICAL TEMPERATURE: 15.6K'

# Assessing correctness in LLMs

Grading LLMs is hard because the answers may be open ended or subject to many formatting issues that make character by character comparison difficult.

There are a couple ways to compare LLM outputs:

- 1) Use multiple choice questions, which makes grading easy. (TruthfulQA, HellaSwag, Winogrande)
- 2) Algorithmic Metrics that programmatically compare output to answers (BLEU, HumanEval)
- 3) Human Evaluation
- 4) Make something up!

# Assessing correctness in LLMs

Task:

- 1) Retrieve the material studied in a given paper
- 2) Find the critical temperature for each material retrieved

Metrics:

- 1) Was the correct number of materials retrieved?
- 2) Was the critical temperature retrieved for each material within 2 K of the database value? (Allows for bad float formatting, also superconducting transition width)

# Cherry-Picked Example

## Model Output:

```
'physrevb.10.4572': ( ['Ga', '69Ga', '71Ga'], [6.1, 6.1, 6.02]),
```

## Database Entry:

```
'physrevb.10.4572': ( ['Ga1'], [6.07]),
```

## ABSTRACT

The superconductive isotope effect in  $\beta$ -phase Ga has been investigated in small single spheres of  $^{69}\text{Ga}$  and  $^{71}\text{Ga}$ . The difference in transition temperatures is  $\Delta T_c = 73 \pm 3 \text{ mK}$ . Assuming a dependence on the mean isotopic mass of the form  $T_c \sim M^{-\alpha}$ , this yields  $\alpha = 0.43 \pm 0.02$ . The transition temperatures are  $T_c^{69} = 6.10 \pm 0.02 \text{ K}$  and  $T_c^{71} = 6.02 \pm 0.02$ . Our previously published value for  $T_c$  in  $\beta$ -Ga is too low owing to an incorrect thermometer calibration. The correct value is  $T_c^\beta = 6.07 \pm 0.03 \text{ K}$  for natural  $\beta$ -Ga.  $H_c(T)$  has been measured both for  $^{69}\text{Ga}$  and  $^{71}\text{Ga}$ . At  $T = 0$ , the difference in critical field is  $H_0^{69} - H_0^{71} = 5 \pm 2 \text{ Oe}$ , in agreement with the similarity principle. Within the experimental accuracy, the deviation of  $H_c(T)$  from a parabola is independent of isotopic mass, yielding  $D(t^2)_{\max} = -0.021 \pm 0.004$ . The slope  $\left(\frac{dH_c}{dT}\right) \Big|_{T_c} = 155 \pm 2 \text{ Oe/K}$ , independent of isotopic mass to this accuracy. These results show  $\beta$ -Ga to be a rather typical weak-coupling superconductor.

Received 16 July 1974

DOI: <https://doi.org/10.1103/PhysRevB.10.4572>

# Cherry-Picked Example

## One example for finding synthesis method:

What is the synthesis method? . . . Several samples of LaIr<sub>2</sub>Si, were prepared from high-purity elements (La, 99.999o purity, Ir, 99.997 purity; Si, 99.999% purity) by arc melting in an argon atmosphere. Melting losses were less than 0.396 in all cases. Homogenization and annealing was performed above 1000'C in an rf induction furnace in vacuo, below 1000degC in sealed quartz capsules under argon. The high-temperature modification was retained by quenching with an argon jet. Differential thermal analysis (DTA) was carried out at temperatures up to 2000degC using W crucibles Powder x-ray difraction (Guinier camera and powder diffractometer, Cu Ka radiation) and metallography served to monitor sample purity and microstructure. . . .

'Based on the provided text, the synthesis method for LaIr<sub>2</sub>Si is described as follows:  
\* Preparation of samples: High-purity elements (La, Ir, and Si) were melted together in an argon atmosphere using an arc furnace. Melting losses were less than 0.4%.  
\* Homogenization and annealing: The samples were heated above 1000°C in an rf induction furnace in vacuo, and below 1000°C in sealed quartz capsules under argon.  
\* Quenching: The high-temperature modification was retained by quenching the samples with an argon jet.  
Note that the text does not provide a specific synthesis method for the high-temperature (LaIr<sub>2</sub>Si-ht) and low-temperature (LaIr<sub>2</sub>Si-lt) modifications, but rather describes the general procedure used to prepare the samples.'

# Lemon-Picked Example

```
{  
'physrevb.100.024508': ([['UI-ThyBe13', 'LaPtGe',  
'PrOsSb12.', 'LaNiC2', 'Re6Zr', 'PrPt4Ge2', 'SrPtAs',  
'UPt3', 'PrOs4Sb12', 'Sr2RuO4', 'PrOs3Sb13'],  
[1.85, 1.85, 1.85, 1.85, 5.85, 1.85, 3.85, 1.85, 0.85,  
1.85, 1.85])  
}  
  
{  
'physrevb.100.024508':  
[['Pr1Pt4Ge12', 'Pr0.9La0.1Pt4Ge12', 'Pr0.7La0.3Pt4Ge12', 'Pr0.5La0.5Pt4Ge1  
2', 'Pr0.3La0.7Pt4Ge12', 'Pr0.2La0.8Pt4Ge12', 'Pr0.1La0.9Pt4Ge12'],  
[7.91, 7.85, 7.87, 7.97, 8.02, 8.07, 8.22])  
}
```

## I. INTRODUCTION

The study of novel superconductors with intrinsic multiple superconducting phases has broadened the understanding of the microscopic origin of unconventional superconductivity (SC) [1,2]. A particular challenge is to clearly identify multiple SC phases in unconventional superconductors and to interpret their SC order parameter(s) (OP).

Gauge symmetry is always broken in the superconducting state. A key indication of multiphase SC is the observation of additional broken symmetry [time-reversal symmetry (TRS), inversion symmetry] at a distinct temperature  $T_m$  below the superconducting transition temperature  $T_c$ . However, superconductors with intrinsic multiple SC phases are extremely rare.  $T_c \sim T_m$  for many superconductors with broken TRS, such as  $\text{Sr}_2\text{RuO}_4$  [3],  $\text{LaNiC}_2$  [4],  $\text{SrPtAs}$  [5], and  $\text{Re}_6\text{Zr}$  [6]. Empirically, superconductors with 4f or 5f electron elements are likely to have complex SC OPs [1,7] that can possibly

TRS [11]. Several thermodynamic experiments show evidence for two SC transitions [12,13] in  $\text{PrOs}_4\text{Sb}_{12}$ , although  $T_m$  is not clearly below the upper  $T_c = 1.85$  K in muon spin relaxation ( $\mu\text{SR}$ ) experiments [11] and it has been argued [14] that the double  $T_c$  nature is not intrinsic. The SC OP of  $\text{PrOs}_4\text{Sb}_{12}$  is a consequence of crystalline-electric-field and strong spin-orbital coupling (SOC) effects [10], and is complicated by the low  $T_c$  and HF behavior.

The isostructural compound  $\text{PrPt}_4\text{Ge}_{12}$ , with a smaller electron effective mass  $m^*$ , is considered to share great similarity with  $\text{PrOs}_4\text{Sb}_{12}$  in the pairing state [15,16], although no multiple SC phases have been reported. The much higher  $T_c \sim 7.8$  K and non-HF state makes  $\text{PrPt}_4\text{Ge}_{12}$  a simpler playground to study SC in Pr-based filled skutterudites. Then the key question is whether the OP in  $\text{PrPt}_4\text{Ge}_{12}$  is complex, including the possibility of a spin-triplet state and broken TRS [16–21].

Failure reason:  
Not enough context  
and  
Information is in a  
table on the 3rd  
page

# Cherry-Picked Examples

Performance for finding the critical temperature given the chemical formula for papers only measuring one sample:

	Maximum Sequence Length	Exact (0.01K)	< 1 K	<2 K
Llama3-8B	4000 characters (~1000 tokens)	34.8%	59.4%	67.9%
	15000 characters (~3000 tokens)	51.9%	74.9%	81.8%
	28000 characters (~7500 tokens)	47.8%	71.6%	78.4%
Llama3-70B	4000 characters (~1000 tokens)	38.1%	65.4%	72.5%
	15000 characters (~3000 tokens)	53.8%	76.9%	82.0%
	28000 characters (~7500 tokens)	54.4%	78.2%	83.4%

# Bad performance of Llama 2

With Llama2, it's harder to coerce the output into a standard format, such as in the following answer:

Based on the information provided in the document, the material studied is:

MATERIAL: GaCNi3

This is mentioned in the document as the composition of the initial ingots and powders used in the experimental synthesis of the material.

## Practical LLMS: Low-Rank Adaptation (LoRA) of neural networks

Want to adapt a (very) large language model to our domain - but we can't afford / don't have enough data to retrain it.

Constrain our weight-updates to the dense layers in fine tuning to be low rank, i.e.

$$W_{i+1} \leftarrow W_i + BA$$
$$W_i \in \mathbb{R}^{n \times m}, B \in \mathbb{R}^{n \times r}, A \in \mathbb{R}^{r \times m}$$
$$r \ll m, n$$

Goal: Fine-tune on existing materials databases where we know the material name and properties.

# Vector Embeddings and Similarity Search

Want to find relevant documents out of a relatively large corpus.

Chunk and embed the sentences of the documents into a high dimensional vector space. Then find the closest vectors (smallest dot products).

$$L = k\text{-argmin}_{i=0:l} \|x - y_i\|_2$$

In principle, we can embed anything we want— including images— to do heterogeneous embedding.

# Practical LLMs: Retrieval Augmented Generation

- In RAG we want our chatbot to have access to a large wealth of information without having to retrain (very expensive) or fine-tune (still expensive).
- Instead we embed our documents into a high dimensional vector space. Embed the question as well, and find the closest vectors. Do the similarity search. Append found sources to the chat to provide context.
- Usually some retraining is done to adapt the model to the RAG task.

[https://doi.org/10.1162/tacl\\_a\\_00530](https://doi.org/10.1162/tacl_a_00530)  
<https://arxiv.org/abs/2312.10997>

# Future Goals

- Improve paragraph and latex extraction (tables!)
- Improve prompting of LLM using prompt scripting: <https://arxiv.org/pdf/2212.06094>
- Fine-tune LLM using all information in the superconductivity database: QLORA (<https://arxiv.org/abs/2305.14314>)
- Extrapolate to all the papers we can get our hands on (arXiv, other sources, etc.)
- Build a database and RAG system to describe them.