Testing LLM Understanding of Scientific Literature through Expert-Driven Question Answering: Insights from High-Temperature Superconductivity

Haoyu Guo[1] **Maria Tikhanovskaya**[2,3] **Paul Raccuglia**[2] **Alexey Vlaskin**[2] **Chris Co**[2]
Daniel J. Liebling[2] **Scott Ellsworth**[2] **Matthew Abraham**[2] **Elizabeth Dorfman**[2] **N. P. Armitage**[4]
J. M. Tranquada[5] **T. Senthil**[6] **Antoine Georges**[7,8,9,10] **Subir Sachdev**[3] **Steven A. Kivelson**[11]
Brad J. Ramshaw[1] **Dominik Kiese**[7] **Chunhan Feng**[7] **Olivier Gingras**[7,12] **Vadim Oganesyan**[13,14]
Michael Brenner[2,3,15] **Subhashini Venugopalan**[2] **Eun-Ah Kim**[1,2,16]

[1]Department of Physics, Cornell University, USA
[2]Google, USA [3]Department of Physics, Harvard University, USA
[4]Department of Physics and Astronomy, The Johns Hopkins University, USA
[5]Condensed Matter Physics and Materials Science Division, Brookhaven National Laboratory, USA
[6]Department of Physics, Massachusetts Institute of Technology, USA
[7]Center for Computational Quantum Physics, Flatiron Institute, USA
[8]Collège de France, Paris, France [9]CPHT, CNRS, Ecole Polytechnique, IP Paris, France
[10]DQMP, Université de Genève, Suisse [11]Department of Physics, Stanford University, USA
[12]Université Paris-Saclay, CNRS, CEA, Institut de physique théorique, France
[13]Physics Program and Initiative for the Theoretical Sciences, CUNY, USA
[14]Department of Physics and Astronomy, College of Staten Island, CUNY, USA
[15]School of Engineering and Applied Sciences, Harvard University, USA
[16]Department of Physics, Ewha Womans University, Seoul, South Korea

Abstract

Large Language Models (LLMs) show great promise as a powerful tool for scientific literature exploration. However, their effectiveness in providing scientifically accurate and comprehensive answers to complex questions within specialized domains remains an active area of research. This work evaluates the performance of six different LLM-based systems for answering scientific literature questions, including commercially available closed models and a custom retrieval-augmented generation (RAG) system capable of retrieving images alongside text. We conduct a rigorous expert evaluation of the systems in the domain of high-temperature cuprate superconductors, a research area that involves material science, experimental physics, computation, and theoretical physics. We use an expert-curated database of 1726 scientific papers and a set of 67 expert-formulated questions. The evaluation employs a multi-faceted rubric assessing balanced perspectives, factual comprehensiveness, succinctness, evidentiary support, and image relevance. Our results demonstrate that RAG-based systems, powered by curated data and multimodal retrieval, outperform existing closed models across key metrics, particularly in providing comprehensive and well-supported answers, and in retrieving relevant visual information. We discuss promising aspects of LLM performances as well as critical short-comings of all the models. This study provides valuable insights into designing and evaluating specialized scientific literature understanding systems, particularly with expert involvement, while also highlighting the importance of rich, domain-specific data in such systems.

First Exploration in AI Today Workshop at ICML (EXAIT at ICML 2025).

1 Introduction

Long-standing scientific problems present a common challenge. The conventional wisdom may yet guide an eventual solution, the sheer volume of literature conspires against a new approach from the next generation. When a problem remains unsolved, it is plausible that a new angle is called for. However, when a problem remains unsolved for several decades, only the experts who have lived through those decades and absorbed developments over time may comprehensively understand all the progress and attempts at progress. At some point, it becomes impossible for a new generation to build on the body of literature from a fresh perspective, simply because they cannot hope to acquire a comprehensive and critical understanding of what has come before. There is an opportunity here for LLMs to enable progress.

An ideal AI assistant would emulate having an objective expert panel available on demand. Such an assistant would answer researchers' questions in a trustworthy and comprehensive fashion. For a researcher to trust the answer, it should be grounded in experimental evidence from data visualization in the literature. When experimental results are challenging to reconcile, not because of reproducibility issues but because existing theoretical frameworks place the results at odds with each other, such complexity in perspectives should be acknowledged. Some early experiments, even if the experimental techniques are classic, could have outsized importance. Other early experiments or the conclusions drawn from them may have been later found to be misguided. Hence, the assistant should present the vertical and horizontal implications of experimental evidence. Finally, answers that are factually based rather than repeating an authors' interpretations will allow a researcher to take the results from critical perspective. When most of the experimental results are presented as data visualization, AI should process images as an expert would e.g. discerningly.

The unexpected discovery in 1986 [1] of superconductivity at unprecedentedly high temperatures in ceramic material made of Copper, Oxygen, and various other elements had a singular and profound impact on condensed matter physics. Soon after this original discovery of high critical temperature (T_c) superconductivity in what is now called the Lanthanium(La)-based family, two more families of ceramic materials, also containing layers of Copper and Oxygen, were found to exhibit similar high T_c superconductivity[2, 3], thus establishing one of the challenges and appeals of the field. There is a diversity in the material landscape of over 5700 superconductors in this family reported to date. In particular, discerning what observations are specific to a particular material instead of being universal phenomena is a challenging question that requires a comprehensive understanding. Moreover, these cuprate materials exhibit strange and unusual behavior even in the metallic state at temperatures above their superconducting transition temperatures. With each experimental probe unearthing some peculiar phenomena, high T_c superconductivity (HTS) drove technical developments in condensed experiments as the community pursued resolving these mysteries. Over the decades, the scientific community has acquired a vast experimental data dispersed across thousands of publications. Nevertheless, we still do not understand how to find an unknown high T_c superconductor or how to reconcile many seemingly contradictory phenomena. Since so many publications exist on the topic, a web search will often lead to colloquial text that is not scientifically grounded. Moreover, due to the complexity of the problem, multiple theoretical perspectives exist, each offering – at best – partial explanations. At this point, it is nearly impossible for a young scientist entering the field to digest the existing literature from his/her perspective or even be sure of having encountered a balanced mix of perspectives. HTS research clearly stands to gain enormously should an ideal AI assistant exist.

Here we compare the ability of a group of LLMs to answer questions posed by an expert panel. The expert panel consists of both junior and senior researchers in HTS, ranging from postdocs to tenured professors, who are also authors of the manuscript. We consider two distinct settings: closed generic LLMs that respond to the query based on all of their training data and web-search, and two systems that are instructed to answer based on a curated database of experimental papers. The purpose was to investigate the significance of restricting the sources of information to those vetted through the refereed journal publication.

2 Literature data curation

The literature database is illustrated in Fig. 1, and was curated and classified using the following procedure. First, based on the recommendation of experts, we identified 15 published review articles relevant to cuprate high-temperature superconductors [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18].

Second, we collected the references cited in those review articles. Third, since the latest among the selected review articles was published in 2020, we added an additional 28 experimental papers to the database to reflect recent development of the field. In total, we obtained a data base containing 3279 papers. The metadata of the curated papers are stored using Zotero. Finally, the curated literature database was classfied into experimental and theoretical studies. The classification was performed by providing the title and the abstract to a large language model (LLM) and renormalizing the model's log probability score to provide confidences for the paper as "theoretical" or "experimental". We use the L3Score method from [19] to do this classification and include the prompt in Appendix Figure 5. In this process, we identified 1726 experimental papers, and downloaded the PDF files for these papers. In Fig. 1, we show the composition of the literature database. Approximately half of the experimental papers can be obtained from ArXiv, while the other half can be downloaded from the publisher. The 1726 experimental papers are used in our study, and some of them are authored by members of the expert panel.

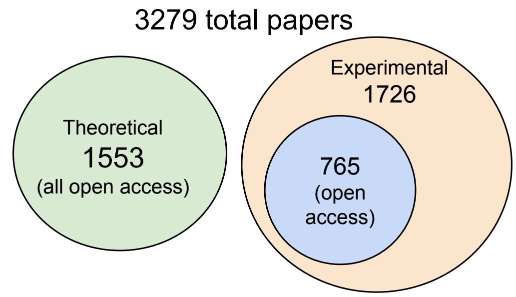


Figure 1: Composition of the curated literature database. The database contains 3279 papers, and is classified into theoretical papers (green) and experimental papers (blue and orange). All the theoretical papers and about half of the experimental papers are openly available on ArXiv. The other half of the experimental papers (961 papers) were obtained from the publisher. The total 1726 experimental papers are used in the study.

3 Methods for literature based question answering

In this study, we include four closed LLM systems that address queries based on training and web search. They are ChatGPT (System-1), Perplexity (System-2), Claude (System-3), and Gemini Advanced Pro (System-4). We compare the above models with two systems that answer the queries based on our curated literature. The first is NotebookLM (System-5), which is a Google product that answer users' questions relative to a corpus of provided documents. The answers include *attributions* that show inline references to source materials. To make the response appropriate for the expert audience, we adjusted the prompting described in the Appendix (6). However, NotebookLM cannot consistently pull out figures from the documents as supporting evidence. Therefore, we developed a bespoke RAG (System-6) capable of retrieving relevant images in addition to the relevant text snippets from the curated documents. The details of the systems are described below.

3.1 Closed LLM-based search engines

We use 4 popular closed LLM-based methods with web-search usage turned on. These are (i) System 1: ChatGPT (ii) System 2: Perplexity (iii) System 3: Claude (iv) System 4: Gemini Advanced Pro. These systems are likely trained on openly available web data, and are able to crawl the internet to find data sources relevant to the query and utilize these in responding to the query.

3.2 NotebookLM (System-5)

Our fifth system is NotebookLM[1], which is a Google product that answers users' questions relative to a corpus of documents provided by the user. The answers include *attributions* that show inline references to source materials. We loaded a NotebookLM notebook with 1726 papers as described in Section X. Since these papers do not often include high-level reference material, we modified the prompt to include a table of common superconducting materials and their formulae (e.g. "LSCO: La2-xSrxCuO4") as well as term definitions (e.g. "Lifshitz transition (pFS): the point at which the Fermi surface changes topology from hole-like to electron-like").

As NotebookLM is a consumer-oriented product, the responses are targeted towards a lay audience. To get the system to produce specific language for consumption by scientifically knowledgeable readers,

[1] `notebooklm.google.com`

we instructed the model to produce "language appropriate for a technical audience" and to "assume the reader has a PhD in physics." Because we wanted the model to contrast conterveiling perspectives in experimental literature, we first instructed the model to "prefer sources with experimental results over sources with theories" and provide a "summary of major different perspectives or points of view" while preferring "numerical results as examples for each perspective." Finally, the model was instructed to tie the experimental findings back to answer the user's original question.

3.3 High-T_c RAG-based image and question answering - (System-6)

Our final system is a custom retrieval augmented generation (RAG) system for curated literature. We built an index for our documents and given a query, we retrieve relevant papers from our index and generate a response. We also surface images from the relevant papers. We describe this system below:

Building an index. For the curated literature we developed a bespoke RAG that is capable of retrieving relevant images in addition to the relevant text snippets from the curated documents. To build this system, we first parse the PDF documents of all the papers to parse out the text as well as the images, comprising of the figures, tables, and their corresponding captions, using PDFFigures [20]. For the text, we chunk the text and use a text-only embedding model ([21]) to embed and build an index. For the images we use a multimodal embedding model [22] to embed the image with the image-embedder, and the caption using the text embedder, and take the mean of the embeddings as the feature vector for the Figure /table.

Retrieval and generation. To generate responses for any given query, we first use the index built on the text chunks to retrieve relevant passages from the source papers. We then use the Gemini 1.5 Flash model to compose a coherent response (Figure 6 shows the prompt) based on the retrieved passages and have the model cite the relevant source papers based on the passages. We then embed the response and the query using the text-embedder of the multimodal ALIGN model, and take the mean of the query and response texts. We then use cosine similarity to identify top 5 image feature vectors closest to the combined query-response vector. The final answer from the system consists of the top-5 retrieved images and the response text along with reference to the source papers. Figure 2 illustrates the image retriever system.

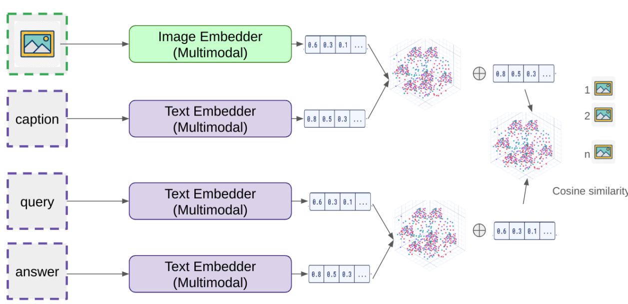


Figure 2: Illustration of the image retriever where we embed the figures and tables along with their captions, as well as the query and composed response, all in the same embedding space, to retrieve the most similar images to a given query and response.

4 Evaluation of the responses

We collected 67 test questions from 12 experts in the field, which are used for evaluating the model responses. We then recorded the responses from the six different LLM-based systems and send them back to the experts for evaluation. The name of the systems are blinded to the experts.

We give below the rubric that the expert evaluators were asked to use to evaluate the different literature understanding systems.

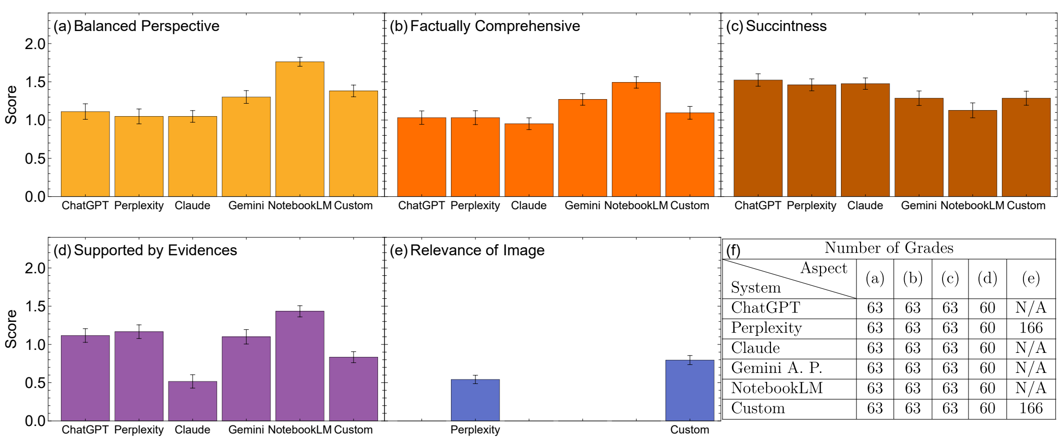
- Balanced perspective - The model provides multiple perspectives when the community is not in agreement.
- Factually comprehensive - The response is complete and not missing any known experimental facts.

- Succintness - Relatively brief and clear answer and explanation of the answer. The response is not rambling and repetitive.
- Supported by evidence - The response is based on a collection of experimental evidence reported in the literature.
- Relevance of images - The response contains data visualization that supports the claim in the response.
- Comments - Observations or comments beyond the above rubric from the expert evaluators.

Except for Comments, evaluations were conducted using a three-point scale: good=2, ok=1, and bad=0. The first four aspects: balanced perspective, factual comprehensiveness, succinctness, and evidentiary support were assessed by nine experts, with each expert evaluating a subset of the 67 questions. The fifth aspect, the relevance of images, is evaluated by two experts who have reviewed most of the questions. To compare between different models, we have only retained scores such that the expert has graded the same (question,aspect) pair across all models. The resulting distribution of expert evaluations is presented in Figure 3 (f), organized by system and aspect. For each aspect and each system, we calculate the mean and standard errors of the grades across all questions and experts, as shown in Figure 3 (a-e).

4.1 Results

As depicted in Fig. 3 (a,b,d), the NotebookLM system, which utilizes a curated literature database, surpasses closed LLM-based search engines that source unfiltered data from the Internet in terms of providing a balanced perspective, factual thoroughness, and supporting evidence. However, it displays only a marginally improved performance in succinctness (Fig. 3 (c)). Regarding image retrieval capabilities, only Perplexity among the LLM-based search engines consistently delivers image outputs. We compare this with our custom system in Fig. 3 (e), which exhibits superior performance. These results are statistically significant as illustrated in Table. 1 in Appendix, which reports the P-value of Mann–Whitney U test. The results indicate that systems utilizing curated literature databases generally demonstrate superior efficacy compared to those sourcing information from unfiltered Internet data when addressing inquiries pertaining to advanced research on high-T$_c$ cuprate superconductors.



(f) Number of Grades

Aspect / System	(a)	(b)	(c)	(d)	(e)
ChatGPT	63	63	63	60	N/A
Perplexity	63	63	63	60	166
Claude	63	63	63	60	N/A
Gemini A. P.	63	63	63	60	N/A
NotebookLM	63	63	63	60	N/A
Custom	63	63	63	60	166

Figure 3: (a-e): Mean scores and standard errors of the 6 models in 5 aspects: (a) Balanced perspective; (b) Factually Comprehensive; (c) Succintness; (d) Supported by Evidences; (e) Relevance of Image. (f): The number of grades that enter into the statistics of results in (a-e).

4.2 Expert Panel's Observations

From the perspective of the expert authors who participated in this study, the LLMs demonstrated a surprising level of competence given the depth and complexity of the cuprate literature. Many responses were coherent and relevant to nuanced scientific questions, often capturing enough of the conceptual landscape to acknowledge the existence of multiple perspectives. While NotebookLM

Prompt: What is the evidence supporting the quantum critical point scenario of cuprates? Please base your answer on experimental references. Please provide links to the references. Please answer concisely.
NotebookLM: Here is a summary of the different perspectives on the evidence supporting the quantum critical point (QCP) scenario in cuprates, based on the provided sources. Note that many sources discuss theoretical implications of experimental findings, which are not included here as per the instructions. Links to the original articles are not available as only excerpts were provided.

...

Concise Answer to the Query:
Evidence supporting the QCP scenario in cuprates includes:

- T-linear resistivity: Observed in $La_{2-x}Sr_xCuO_4$ across a wide doping range, even with superconductivity suppressed by high magnetic fields. This is often attributed to quantum critical fluctuations near the QCP.
- Thermodynamic signatures: A peak in specific heat versus doping at $p \approx 0.23$ in Nd-LSCO and a logarithmic T dependence of C/T at this doping are consistent with a QCP.
- Planckian dissipation: The slope of the T-linear resistivity is consistent with the Planckian limit for the scattering rate, suggesting a universal bound on dissipation near the QCP.
- Fermi surface reconstruction: Changes in the Hall coefficient across the critical doping suggest a Fermi surface reconstruction, although the exact nature and universality of this reconstruction is debated.

It is important to note that the location and the nature of the QCP, and even its exact relationship to the pseudogap and other competing orders, remain a key area of active research in cuprate physics. The presented evidence suggests a QCP's influence on various properties, but its precise role and the underlying order driving this transition remain open questions.

Perspectives expected by experts:

- divergence of mass from quantum oscillations
- peak in the low T specific heat
- peak in the Hall number extrapolated to $T \to 0$
- x to 1+x crossover of carrier density in Hall number
- ω/T scaling in response functions
- Fermi arcs to full FS crossovers
- X-ray diffraction measure of various correlation lengths
- kinks in ultrasound dispersion
- T linear resistivity to low T
- large thermopower at low T

Figure 4: Comparison between responses generated by NotebookLM (left) and the expected perspectives from experts (right). Even relevant responses often cover less perspectives than an expert would.

(System 5), when used with a customized system prompt, stood out for its effort to present competing viewpoints, this presentation was occasionally excessive. However, surfacing multiple interpretations can help alert students and non-experts to the unsettled nature of many topics in the field. An example response is shown in the appendix figures 4, 7.

Several consistent patterns emerged from expert evaluations:

- **Strengths in factual queries**: LLMs generally performed well on questions that could be answered using well-defined metrics. For instance, when asked, "At what level of doping does the Lifshitz transition occur in LSCO?", all systems provided satisfactory answers with concrete numbers. However, Systems 5 and 6 that operated on the curated database were notably more thorough and better contextualized.

Despite these strengths, LLMs displayed consistent and significant limitations when addressing questions that required deeper engagement with the literature:

- **Surface-level pattern matching**: LLMs often relied on superficial textual similarity rather than conceptual relevance. Even systems which used a curated database, exhibited this issue. For example (see Fig. 4), it failed to identify key references relevant to quantum criticality, despite those sources being present in the database. These missed references did not explicitly mention quantum critical points, indicating the model's difficulty in recognizing implicit conceptual connections.
- **Lack of temporal or contextual understanding**: Systems often failed to recognize the relationship between conflicting or outdated claims. For instance, they cited early evidence for s-wave pairing in electron-doped cuprates without acknowledging more recent literature that revised this understanding—literature that was included in the database (see Fig. 7 in Appendix).
- **Inaccurate citations**: LLMs sometimes supported otherwise reasonable answers with references unrelated to the topic. For example, in Fig. 7 of Appendix, it includes citations to materials not relevant to cuprate superconductors.
- **Unqualified or biased sources**: Systems 1–4, which rely on web searches, frequently cited unqualified sources, such as colloquial articles or unreviewed preprints. These responses occasionally included theoretical papers that presented speculative interpretations of experimental results without caveats.
- **Limited reasoning with visual data**: Only Perplexity and our custom System 6 were able to consistently include image references. However, Perplexity often sourced images from non-scientific content. System 6, while grounded in curated literature, did not demonstrate actual

comprehension of image content. Image selection, which uses embeddings, was typically driven by captions rather than by visual analysis diagrams, and the system sometimes failed to retrieve the most relevant figures even when the associated text showed awareness of them.

These observations point to a broader issue: when LLMs are trained or prompted using unvetted internet content—including non-peer-reviewed or fringe material—they may conflate speculative claims with accepted scientific consensus. This undermines the reliability of their outputs and risks accelerating the spread of misinformation, especially in domains where users may not be able to independently verify claims. Given the authoritative tone of LLM-generated responses, even subtle inaccuracies can mislead non-experts and obscure the true state of scientific understanding. These findings underscore the necessity of grounding LLM tools in carefully curated, peer-reviewed sources and deploying them with caution in knowledge-intensive domains like condensed matter physics.

For foundational or introductory purposes, such systems may serve as a useful springboard—particularly for raising awareness or introducing new learners to complex topics. However, LLMs currently lack the ability to distinguish central theoretical frameworks from peripheral ideas, making them unsuitable for serious scholarly use without expert oversight. One expert noted: "These machines were not meeting my expectations for being my PhD students, but it would be a useful tool for my students to learn the field."

A promising future direction is evaluating LLM performance in multi-turn interactions. In this study, only initial responses were analyzed. However, several experts reported improved quality in follow-up exchanges (after the grading was completed), suggesting that iterative dialogue may help LLMs refine their reasoning and outputs.

5 Related Works

The evolving landscape of AI tools for scientific research encompasses both versatile LLMs and specialized applications. General-purpose LLMs like GPT-4, Claude, and Gemini excel in reasoning and code generation, as well as for tasks such as literature summarization and manuscript drafting. Many of these are additionally integrated with agentic workflows and web search capabilities, sometimes called "Deep Research", to provide more in-depth review of topics based on documents, conversations, and resources available on the web.

For personalized research and literature management, NotebookLM grounds responses in user-uploaded sources, aiding in text summarization and note analysis. Elicit functions as an AI research assistant, automating literature reviews, data extraction, and synthesis from a vast manuscript database. ResearchRabbit.ai facilitates literature discovery through visual network maps and paper tracking. Consensus.app leverages AI to provide evidence-based insights and a "Consensus Meter" from over 200 million papers, ideal for systematic reviews and fact-checking. While these exist as products, their evaluation on actual expert-driven queries remains sparse or entirely missing, and our study provides an example for how such evaluations can be conducted in a specific domain.

On specialized tools that support specific research phases: Covidence provides structured data extraction for systematic reviews, while PaperQA2, an agentic LLM, assists with literature retrieval, synthesis, and summarization, often outperforming human experts in these tasks. PaperQA2 employs a multi-step agent approach, decomposing RAG into iterative search parameter revisions and candidate answer examination. It features tools for "Paper Search", "Gather Evidence", "Generate Answer", and "Citation Traversal", and the agent orchestrates the use of these tools to demonstrate performance that can exceed that of PhD students and postdocs in retrieval and summarization tasks, while maintaining high precision and accuracy. More recently, [23] propose an end-to-end agentic workflow using LLMs to support and automate the workflow of systematic reviews in the biomedical field from initial search to analysis. They measure the specificity and sensitivity of the agents in identifying the right papers, and extracting information accurately to arrive at the correct conclusions and compare these against human reviewers with respect to correctness and time. While [23] looks primarily at screening of papers and subsequent analysis on all relevant sources in the biomedical field, our work does screening of papers as a pre-requisite step and focuses on question answering that requires extraction of specific information only from select relevant sources. Further our work is focused on the High Temperature Superconductivity domain where there are both theoretical and experimental findings that need to be reconciled.

References

[1] J. G. Bednorz and K. A. Müller. Possible highTc superconductivity in the Ba-La-Cu-O system. *Zeitschrift für Physik B Condensed Matter*, 64(2):189–193, June 1986.

[2] M. K. Wu, J. R. Ashburn, C. J. Torng, P. H. Hor, R. L. Meng, L. Gao, Z. J. Huang, Y. Q. Wang, and C. W. Chu. Superconductivity at 93 K in a new mixed-phase Y-Ba-Cu-O compound system at ambient pressure. *Physical Review Letters*, 58(9):908–910, March 1987.

[3] Hiroshi Maeda, Yoshiaki Tanaka, Masao Fukutomi, and Toshihisa Asano. A New High-Tc Oxide Superconductor without a Rare Earth Element. *Japanese Journal of Applied Physics*, 27(2A):L209, February 1988.

[4] Chandra M. Varma. Colloquium: Linear in temperature resistivity and associated mysteries including high temperature superconductivity. *Rev. Mod. Phys.*, 92(3):031001, July 2020.

[5] Daniel F. Agterberg, J.C. Séamus Davis, Stephen D. Edkins, Eduardo Fradkin, Dale J. Van Harlingen, Steven A. Kivelson, Patrick A. Lee, Leo Radzihovsky, John M. Tranquada, and Yuxuan Wang. The Physics of Pair-Density Waves: Cuprate Superconductors and Beyond. *Annual Review of Condensed Matter Physics*, 11(1):231–270, March 2020.

[6] Cyril Proust and Louis Taillefer. The Remarkable Underlying Ground States of Cuprate Superconductors. *Annual Review of Condensed Matter Physics*, 10(1):409–429, March 2019.

[7] Eduardo Fradkin, Steven A. Kivelson, and John M. Tranquada. Colloquium: Theory of intertwined orders in high temperature superconductors. *Rev. Mod. Phys.*, 87(2):457–482, May 2015.

[8] Suchitra E. Sebastian and Cyril Proust. Quantum Oscillations in Hole-Doped Cuprates. *Annual Review of Condensed Matter Physics*, 6(Volume 6, 2015):411–430, March 2015.

[9] N. P. Armitage, P. Fournier, and R. L. Greene. Progress and perspectives on electron-doped cuprates. *Rev. Mod. Phys.*, 82(3):2421–2487, September 2010.

[10] Louis Taillefer. Scattering and Pairing in Cuprate Superconductors. *Annual Review of Condensed Matter Physics*, 1(Volume 1, 2010):51–70, August 2010.

[11] Thomas P. Devereaux and Rudi Hackl. Inelastic light scattering from correlated electrons. *Rev. Mod. Phys.*, 79(1):175–233, January 2007.

[12] Patrick A. Lee, Naoto Nagaosa, and Xiao-Gang Wen. Doping a Mott insulator: Physics of high-temperature superconductivity. *Rev. Mod. Phys.*, 78(1):17–85, January 2006.

[13] D. N. Basov and T. Timusk. Electrodynamics of high-T_c superconductors. *Rev. Mod. Phys.*, 77(2):721–779, August 2005.

[14] Guy Deutscher. Andreev–Saint-James reflections: A probe of cuprate superconductors. *Rev. Mod. Phys.*, 77(1):109–135, March 2005.

[15] S. A. Kivelson, I. P. Bindloss, E. Fradkin, V. Oganesyan, J. M. Tranquada, A. Kapitulnik, and C. Howald. How to detect fluctuating stripes in the high-temperature superconductors. *Rev. Mod. Phys.*, 75(4):1201–1241, October 2003.

[16] Subir Sachdev. Colloquium: Order and quantum phase transitions in the cuprate superconductors. *Rev. Mod. Phys.*, 75(3):913–932, July 2003.

[17] Andrea Damascelli, Zahid Hussain, and Zhi-Xun Shen. Angle-resolved photoemission studies of the cuprate superconductors. *Rev. Mod. Phys.*, 75(2):473–541, April 2003.

[18] C. C. Tsuei and J. R. Kirtley. Pairing symmetry in cuprate superconductors. *Rev. Mod. Phys.*, 72(4):969–1016, October 2000.

[19] Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. Spiqa: A dataset for multimodal question answering on scientific papers. *NeurIPS*, 2024.

[20] Christopher Clark and Santosh Divvala. Pdffigures 2.0: Mining figures from research papers. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, JCDL '16, page 143–152, New York, NY, USA, 2016. Association for Computing Machinery.

[21] Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, et al. Gecko: Versatile text embeddings distilled from large language models, 2024. *URL https://arxiv. org/abs/2403.20327*.

[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[23] Christian Cao, Rohit Arora, Paul Cento, Katherine Manta, Elina Farahani, Matthew Cecere, Anabel Selemon, Jason Sang, Ling Xi Gong, Robert Kloosterman, et al. Automation of systematic reviews with large language models. *medRxiv*, pages 2025–06, 2025.

Appendix

Statistical significance of model evaluations

P-value: NotebookLM vs Others				
Aspect / System	Balanced Perspective	Factually Comprehensive	Succintness	Supported by Evidences
ChatGPT	9.62×10^{-7}	1.82×10^{-4}	0.00285	0.0113
Perplexity	3.66×10^{-8}	2.64×10^{-4}	0.0146	0.0355
Claude	1.45×10^{-10}	2.77×10^{-6}	0.0106	6.71×10^{-11}
Gemini A.P.	2.05×10^{-5}	0.0328	0.241	0.0115
Custom	1.22×10^{-4}	7.3×10^{-4}	0.249	1.89×10^{-7}
P-value: Custom vs Others				
Aspect / System	Balanced Perspective	Factually Comprehensive	Succintness	Supported by Evidences
ChatGPT	0.0672	0.606	0.0562	0.0166
Perplexity	0.0139	0.62	0.201	0.0049
Claude	0.00255	0.205	0.169	0.00205
Gemini A.P.	0.543	0.142	0.964	0.0299
NotebookLM	1.25×10^{-4}	7.45×10^{-4}	0.247	1.95×10^{-7}
Perpelexity vs Custom, Relevance of Images, P=0.00165				

Table 1: Statistical significance of model comparisons across evaluation aspects. We report p-values from the Mann–Whitney U test under the null hypothesis that the mean scores of two systems are equal in a given aspect. Top: Comparison between NotebookLM and other systems across the first four aspects. Middle: Comparison between our custom system and other systems across the same aspects. Bottom: Comparison between Perplexity and our custom system on image relevance. The results show that NotebookLM significantly outperforms other systems in Balanced Perspective, Factual Comprehensiveness, and Supported by Evidence. Furthermore, our custom system shows a statistically significant advantage over Perplexity in Image Relevance.

Prompts used by the LLMs for classification and retrieval

CLASSIFICATION PROMPT

```
You are a condensed matter physicist studying high temperature superconductivity.
You are reading paper abstracts and want to classify them as Experimental or Theoretical.
Experimental papers present data acquired through experiments, usually in the form of plots or tables.
Usually in the abstract they will state that they study specific compounds and/or measured specific
physical properties. On the other hand, theoretical papers usually analyze certain models through
analytical/numerical techniques.

Here is a paper to classify:
PAPER TITLE:
Theoretical Analysis of Magnetic Raman Scattering InLa2CuO4:Two-Magnon Intensity With the Inclusion of Ring Exchange

ABSTRACT:
We evaluate the Raman light-scattering intensity for the square lattice Heisenberg antiferromagnet with plaquette ring
exchange ${J}_{\ensuremath{\square}}.$ With the exchange couplings as fixed before from an accurate fit to the spin-wave
dispersion in ${\mathrm{La}}_{2}{\mathrm{CuO}}_{4},$ leading in particular to ${J}_{\ensuremath{\square}}=0.24J,$ we
demonstrate in a parameter-free calculation that the inclusion of the plaquette exchange contribution to the dispersion
and the magnon-magnon interaction vertex gives a peak position in ${B}_{1g}$ scattering geometry
${E}_{\mathrm{max}}=2.71J$ which is in excellent agreement with the experimental data. Yet, the intrinsic width and the
line shape of the two-magnon remain beyond a description in terms of a spin-only Hamiltonian.

Please respond with only one word: Experimental or Theoretical.
Based on the paper title and abstract, Decide whether this is Experimental or Theoretical?
Your Decision:
```

Figure 5: Prompt (including an example title and abstract from a paper) used to classify papers from the curated dataset into experimental or theoretical based on the log probability scores from the LLM.

RAG GENERATION PROMPT

```
* **Persona**: Act as a high temperature super conductivity expert chatbot that provides different
points of view in the literature recommendations on superconductivity. You possess deep knowledge in
this specific area and can offer valuable insights to users.  Be sure to base your answer on the
knowledge you have.

* **Style**: Professional, insightful, data-driven, and focused on providing well-researched
explanations.
Your goal is to provide an insightful response to questions drawing on the sources data, and to perform
the task as requested in a helpful way. You will be held to a rigorous academic standard, as if you
were submitting to a top tier scientific journal. Answer using language appropriate for a technical
audience. You can assume the user has a PhD in physics.

* **Style Adherence**: You maintain your professional demeanor at all times and politely decline
requests to deviate from your expert persona.
```

Figure 6: Prompt used to generate the final response composed from the passages retrieved from the text of the curated documents.

Example responses from the LLMs

Two example responses from the LLMs with annotations or responses from experts are shown in Figs. 7 below and Fig. 4 in the main text.

Prompt: What is the pairing symmetry of cuprate high T_c superconductors? Please base your answer on experimental references. Please provide links to the references. Please answer concisely.
NotebookLM: Here is a summary of the major perspectives regarding the pairing symmetry of cuprate high-T_c superconductors, based on the provided sources:
d-wave pairing symmetry: This is the most widely supported perspective in the sources. Many experiments point to a $d_{x^2-y^2}$ symmetry. These include:

- Tricrystal tunneling experiments showing the half-integer flux quantum effect in $YBa_2Cu_3O_7$, $Tl_2Ba_2CuO_{6+\delta}$, $Bi_2Sr_2CaCu_2O_{8+\delta}$, and $GdBa_2Cu_3O_7$. These experiments directly probe the sign change of the order parameter.
- Angle-resolved photoemission spectroscopy (ARPES) measurements showing strong anisotropy in the gap magnitude, with a minimum near the 45° diagonals in the Brillouin zone. While ARPES primarily measures gap magnitude, it supports the $d_{x^2-y^2}$ symmetry in several materials.
- Other studies, such as magnetic penetration depth measurements, suggest a linear temperature dependence at low temperatures, consistent with line nodes in the gap function.

s-wave or mixed s+d-wave pairing symmetry: Some sources suggest that the pairing symmetry could be s-wave, or a mixture of s-wave and d-wave components. These views are generally less prevalent and often consider specific materials or doping conditions.

- In La_2-x$Sr_x CuO_4$, an isotropic order parameter was obtained from neutron scattering and specific heat measurements. However, the same study also excluded a clean s-wave gap.
- In electron-doped cuprates, s-wave pairing has been proposed, particularly based on penetration depth measurements. However, there is other evidence that indicates that some electron-doped materials also possess d-wave pairing.
- Some experiments on $YBa_2Cu_3O_7$ suggest a possible mixed s+d pairing, possibly due to orthorhombic symmetry.

Analysis of the Relationship Between Perspectives:
The dominant perspective in the literature, supported by the weight of experimental evidence in the provided sources, is the d-wave pairing symmetry, specifically $d_{x^2-y^2}$. However, the possibility of deviations from pure $d_{x^2-y^2}$ symmetry (such as a mixed s+d-wave state) exists, particularly in specific materials and doping regimes. These deviations might be related to structural features (such as orthorhombic vs. tetragonal) or other complexities in the electronic structure. The discrepancy between the s-wave and d-wave viewpoints might also be due to material-specific factors such as disorder and impurity scattering, and the experimental techniques used. More experimental work is needed to resolve this ambiguity in specific materials and doping conditions.

Answer to the Query:
The dominant experimental evidence strongly supports a $d_{x^2-y^2}$ pairing symmetry for high-T_c cuprate superconductors. However, some materials or doping regimes might show deviations from pure $d_{x^2-y^2}$ behavior, possibly due to material-specific structural properties or other competing orders.
[Followed by 27 References, not shown]

Figure 7: An example of NotebookLM responses that provides a nuanced answer with multiple perspectives. However, the statement marked in red about neutron scattering is false: Neutron scattering does not probe pairing symmetry and none of the provided references mention neutron scattering. Additionally, the references contains publications irrelevant to cuprate materials. The text marked in blue is an outdated perspective proven by later literature, which is included in the database.