PNAS RESEARCH ARTICLE | PHYSICS OPEN ACCESS

Expert evaluation of LLM world models: A high-T_c superconductivity case study

Haoyu Guo[a,1], Maria Tikhanovskaya[b,c,1], Paul Raccuglia[b], Alexey Vlaskin[b], Chris Co[b], Daniel J. Liebling[b], Scott Ellsworth[b], Matthew Abraham[b], Elizabeth Dorfman[b], N. P. Armitage[d], Chunhan Feng[e], Antoine Georges[e,f,g,h], Olivier Gingras[e,i], Dominik Kiese[e], Steven A. Kivelson[j], Vadim Oganesyan[k,l], B. J. Ramshaw[a], Subir Sachdev[c,2], T. Senthil[m], J. M. Tranquada[n], Michael P. Brenner[b,c,o], Subhashini Venugopalan[b], and Eun-Ah Kim[a,b,p,2]

Affiliations are included on p. 8.

Contributed by Subir Sachdev; received November 26, 2025; accepted January 30, 2026; reviewed by Maissam Barkeshli, Marc-Henri Julien, and Yi-Zhuang You

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. Using the field of high-temperature cuprates as an exemplar, we evaluate the ability of LLM systems to understand the literature at the level of an expert. We construct an expert-curated database of 1,726 scientific papers that covers the history of the field, and a set of 67 expert-formulated questions that probe deep understanding of the literature. We then evaluate six different LLM-based systems for answering these questions, including both commercially available closed models and a custom retrieval-augmented generation (RAG) system capable of retrieving images alongside text. Experts then evaluate the answers of these systems against a rubric that assesses balanced perspectives, factual comprehensiveness, succinctness, and evidentiary support. Among the six systems, two using RAG on curated literature outperformed existing closed models across key metrics, particularly in providing comprehensive and well-supported answers. We discuss promising aspects of LLM performances as well as critical short-comings of all the models. The set of expert-formulated questions and the rubric will be valuable for assessing expert level performance of LLM based reasoning systems.

retrieval augmented generation | LLM | cuprate | superconductor

There is a structural problem that impedes progress on long-standing scientific problems; while the body of accumulated wisdom in the field contains valuable information, the sheer volume of literature makes exploiting this knowledge base extremely difficult for a new generation, even when they come armed with potentially game-changing insights, methodologies, or new information. When a problem has remained interesting but only partially solved despite decades of work, only experts with long engagement with the field may appreciate what has been established, including reasons for abandoning seemingly promising lines of thought. At some point, it becomes impossible for a new generation to build on the body of literature from a fresh perspective, simply because it is difficult to acquire a comprehensive and critical understanding of what has come before. There is an opportunity here for LLMs to enable progress.

Ideally, a curious student would have an objective expert panel available on demand, answering researchers' questions in a trustworthy and comprehensive fashion. The goal of this paper is to evaluate whether Large Language Model (LLM) guided assistants can accomplish this goal. Within the experimental sciences, for a researcher to trust an answer, it must be grounded in experimental evidence, including data visualization of the relevant figures in the literature. When experimental results are challenging to reconcile, not because of reproducibility but because existing theoretical frameworks place the results at odds with each other, the complexity in perspectives should be acknowledged. For expert researchers, early experiments can have outsided importance, even if the experimental techniques are classic. Other early experiments or the conclusions drawn from them may have been later found to be misguided. Hence, the assistant should present the implications of the experiment in the context of its timing, specifically how it supports or counters previous observations and whether it is in harmony with other contemporary observations, as determined from a known theoretical framework. Perhaps the most valuable quality to seek in an ideal assistant is sound and critical judgment,

Significance

Solving long-standing scientific problems requires researchers to navigate vast and complex research literatures with competing perspectives. While Large Language Models (LLMs) can aid in this challenging process in principle, reading scientific research literature to synthesize knowledge and identify conclusions supported by experimental evidence requires skills that go beyond everyday reading. Using high-temperature superconductivity as an exemplar, we construct a unique dataset of questions and answers that probe expert knowledge of the literature. We then carry out systematic analysis of multiple LLM systems for answering questions at the level of an expert researcher. We find that curated data and multimodal retrieval significantly improve accuracy and depth. These findings inform the development of trustworthy AI tools to advance scientific progress.

Competing interest statement: The authors affiliated with Google are employees of Google Inc. The authors affiliated with Google hold Alphabet stock. NotebookLM and Gemini used in the study are products of Google.

Copyright © 2026 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1]H.G. and M.T contributed equally to this work.

[2]To whom correspondence may be addressed. Email: sachdev@g.harvard.edu or eun-ah.kim@cornell.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2533676123/-/DCSupplemental.

Published March 10, 2026.

Downloaded from https://www.pnas.org by Ewha Womans University on September 28, 2026 from IP address 203.255.190.49.

enabling them to see beyond the author's bias and interpretation to extract objective facts.

Here, we evaluate the ability of LLM systems to serve as such an assistant in a particular problem of great scientific and technological importance: high-temperature superconductivity (HTS). The unexpected discovery in 1986 (1) of superconductivity at unprecedentedly high-temperatures in ceramic materials 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-temperature (high-T_c) superconductivity in what is now called the 214 (La_2CuO_4-based) family, additional structural families—most notably Y-Ba-Cu-O (YBCO) and the Bi-Sr-Ca-Cu-O (BSCCO) series—were also found to exhibit high-T_c superconductivity (2, 3). Many further families were identified in subsequent years, including Tl- and Hg-based cuprates as well as the electron-doped compounds (4–6), thus establishing both the challenges and appeals of the field. There is a diversity in the material landscape as each family of materials exhibits a rich set of phenomena upon changes in temperature, magnetic field, crystal structure, and charge carrier concentration. Moreover, these cuprate materials exhibit strange and unusual behavior even in the metallic state at temperatures above their superconducting transition temperatures. There are quantitative and qualitative variations in observed phenomena in this high-dimensional parameter space of materials explored by a large community using various samples over the span of four decades. However, discerning what observations are specific to a particular sample, a particular subclass of materials, or a particular family, rather than being universal phenomena, is challenging without a comprehensive understanding of the literature.

Over the decades, the scientific community has amassed a vast body of experimental data, dispersed across thousands of publications. Nevertheless, we still do not understand how to find a new high-T_c superconductor or how to reconcile the many seemingly contradictory phenomena observed in this material class. The long-standing puzzles invited an ever-increasing list of experimental probes to tackle the problem from different angles, each revealing new facets of the problem: angle-resolved photoemission spectroscopy (ARPES), scanning tunneling microscopy (STM), neutron scattering, resistivity, NMR (NMR), nuclear quadrupolar resonance (NQR), THz spectroscopy, optical conductivity, thermal transport, specific heat, muon spin rotation (μSR), X-ray scattering, electron energy loss spectroscopy (EELS), raman scattering, magnetization, angle-dependent magnetoresistance (ADMR), and more. A satisfactory theory must synthesize and reconcile experimental facts gathered from many complementary probes rather than relying on any single measurement. However, it is punishingly challenging for a new researcher to acquire a comprehensive and critical understanding of what has come before, not to mention the challenge of synthesizing disparate results. Moreover, due to the complexity of the problem, multiple theoretical perspectives often 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 their perspective or even be sure of having encountered a balanced mix of perspectives. HTS research would stand to gain enormously if an ideal AI assistant existed.

How close are current LLM systems to achieving these goals? To evaluate this, we use an expert panel to develop a set of materials about HTS that makes it possible to measure the performance of a set of LLM systems against our requirements. The panel designed a set of questions and answers to probe a deep understanding of this literature. Also, it developed a set of papers that defines the scientific literature in this specialized field. The literature in this field is large but finite: The experts selected a set of 1,726 scientific papers that cover the field's history exclusively through experimental observations and discoveries. To probe a deep understanding of this literature, the experts formulated a set of 67 questions and answers covering all aspects of the field, from experimental measurements that define the phenomenon to theoretical ideas that purport to explain them. While some questions have widely agreed-upon answers, others are more nuanced, with differing perspectives or conflicting experimental measurements. The questions aim to measure the ability of the LLM system to appreciate the nuances.

We evaluate the ability of LLM systems to answer these questions in two distinct settings: The first uses closed generic LLMs that respond to the query based on all of their training data and web-search. The second gives the LLM the entire curated database of experimental papers, and asks the system to ground the responses within this literature. We test 6 different LLM systems for answering these questions, including both commercially available closed models and a custom retrieval-augmented generation (RAG) system capable of retrieving images alongside text.

To evaluate performance, the expert panel then manually grades the answers of these systems against a rubric that assesses balanced perspectives, factual comprehensiveness, succinctness, and evidentiary support. While grading, each expert is blinded to the identities of the different LLM systems that produce the answers. By comparing against different types of LLM systems, we not only allow measuring the ability of current AI systems to act as expert assistants, but also measure the significance of restricting the sources of information to those vetted through the refereed journal publication. We note that measuring the performance in this manner is highly laborious and expensive—the complexity of the task means that accurate evaluation can only be done with world experts with deep experience in the field. The uniqueness of the present study lies in putting these pieces together, giving an accurate snapshot of how far LLM technology is from being the ideal AI assistant we seek.

Literature Data Curation

The first step in this process is to curate a complete literature database for HTS (Fig. 1). We curated the database as follows: First, based on the recommendation of experts, we identified 15 published review articles relevant to cuprate high-temperature superconductors (7–21). 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, this leads to set of 3,279 papers. The metadata of the curated papers were stored using Zotero. Finally, the curated literature database was classified into experimental and theoretical studies. We did this by providing the title and the abstract of each paper to a large language model (LLM) and renormalizing the model's log probability score to provide confidence scores for the paper as "theoretical" or "experimental." We used the L3Score method from ref. 22 to do this classification (*SI Appendix*, Fig. S3). This identified 1,726 of the 3,279 papers as experimental, and used these as the primary sources for this study. These experimental papers were then downloaded at Cornell University into a

Downloaded from https://www.pnas.org by Ewha Womans University on September 28, 2026 from IP address 203.255.190.49.

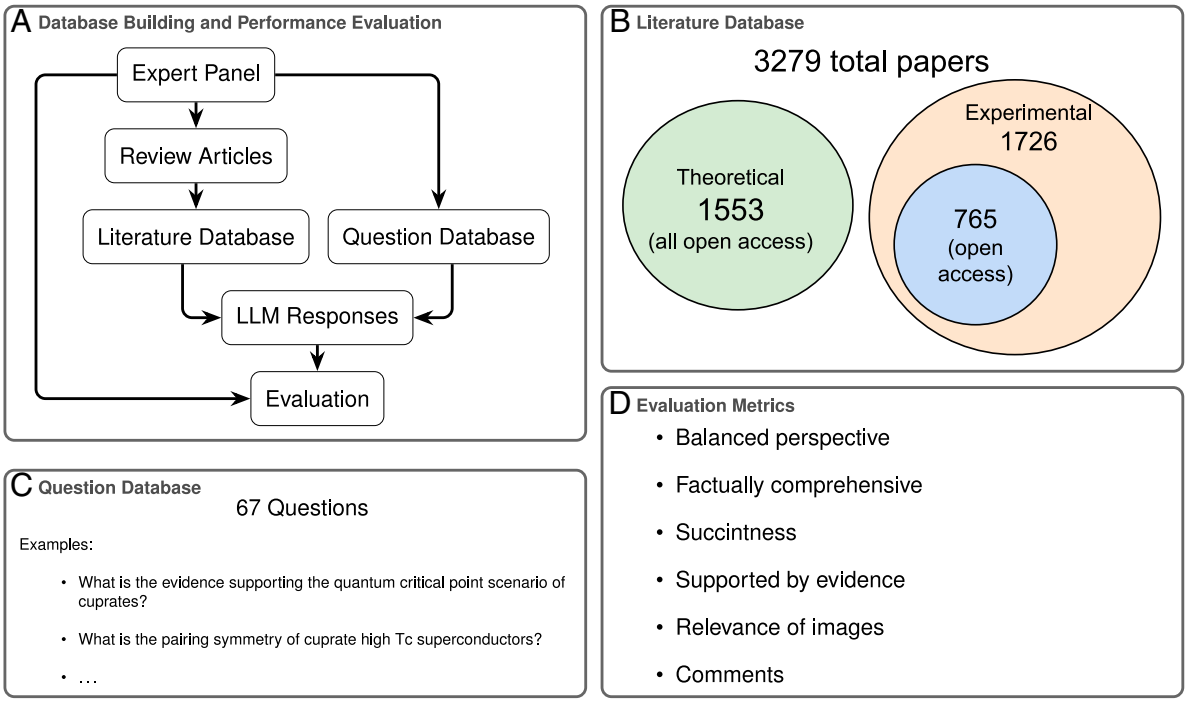


Fig. 1. (*A*) Flow diagram showing the database building process and how the LLMs are evaluated. We curated a literature database based on references of review articles recommended by the expert panel. We also collected questions related to the topic of high-T_c cuprates from the expert panel. The LLMs were prompted to answer these questions and the outputs were graded by the expert panel. (*B*) Composition of the curated literature database. The database contains 3,279 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. A total of 1,726 experimental papers are used in the study. (*C*) Examples of the question database. (*D*) The metrics that the expert panel used to evaluate the LLM outputs.

Downloaded from https://www.pnas.org by Ewha Womans University on September 28, 2026 from IP address 203.255.190.49.

private repository. Fig. 1 shows the composition of the literature database. Approximately half of the experimental papers can be obtained from arXiv.

Question and Answer Dataset

The expert panel consisted of 12 experts with wide ranging knowledge and experience in the field of high-temperature superconductivity. To probe the ability of an LLM system to respond like an expert, the panel then constructed a set of questions and answers that probe deep knowledge of the field. The questions were designed to get at the essential questions underlying research in high-temperature superconductivity and probe deep understanding of the literature. Each expert has differing perspectives on the most important aspects of the field, and the goal is to create questions and answers that broadly cover areas and these perspectives. Overall, we collected 67 test questions.

The questions they constructed delve into various aspects of the cuprate high-T_c superconductor materials, covering their fundamental properties and complex behaviors across different phases. Key areas of inquiry include the evidence, characteristics, and the role of quantum critical points, the nature of charge carriers and the dependence of their density on doping, and the angular and temperature dependencies of transport scattering rates, as well as probing quantitative and qualitative specifics of superconductivity and related phenomena in cuprates. They include investigations into the pairing symmetry of high-Tc superconductors, the evidence for symmetry breaking in the normal state, and the size and properties of vortices. Questions also probe characteristics of various other orders, such as the incommensurability of spin fluctuations, the driving forces behind stripe order (spin or charge), the energy scales associated with potential bosonic fluctuations that mediate pairing, as well as more exotic phenomena and the validity of different theoretical frameworks. In Fig. 2*A*, we categorize the questions in terms of the physical concepts involved in the questions and types of relevant measurement techniques.

The question database reflects the complexity of scientific inquiry in cuprate high-temperature superconductors, a canonical example of a material whose properties stem from complex quantum many-body physics. To address each question successfully, one must invoke multiple theoretical concepts and synthesize accumulated information from measurements using multiple experimental probes. One such example is shown in Fig. 2. For the example question in Fig. 2*B*, the expert panel identified ten distinct aspects that should be included in a comprehensive answer. Fig. 2 *C* and *D* present excerpts from the best-performing LLM responses for text and image, respectively, which nevertheless cover only a small subset of these ten facets.

The expert panel created a rubric to assess a reasoning system's ability to understand the literature. This rubric includes the following elements:

Balanced perspective - The model provides multiple perspectives when the community is not in agreement. A good response should alert the reader to different viewpoints on the queried topic in the literature.

Factually comprehensive - The response is complete and not missing any known experimental facts. A good response

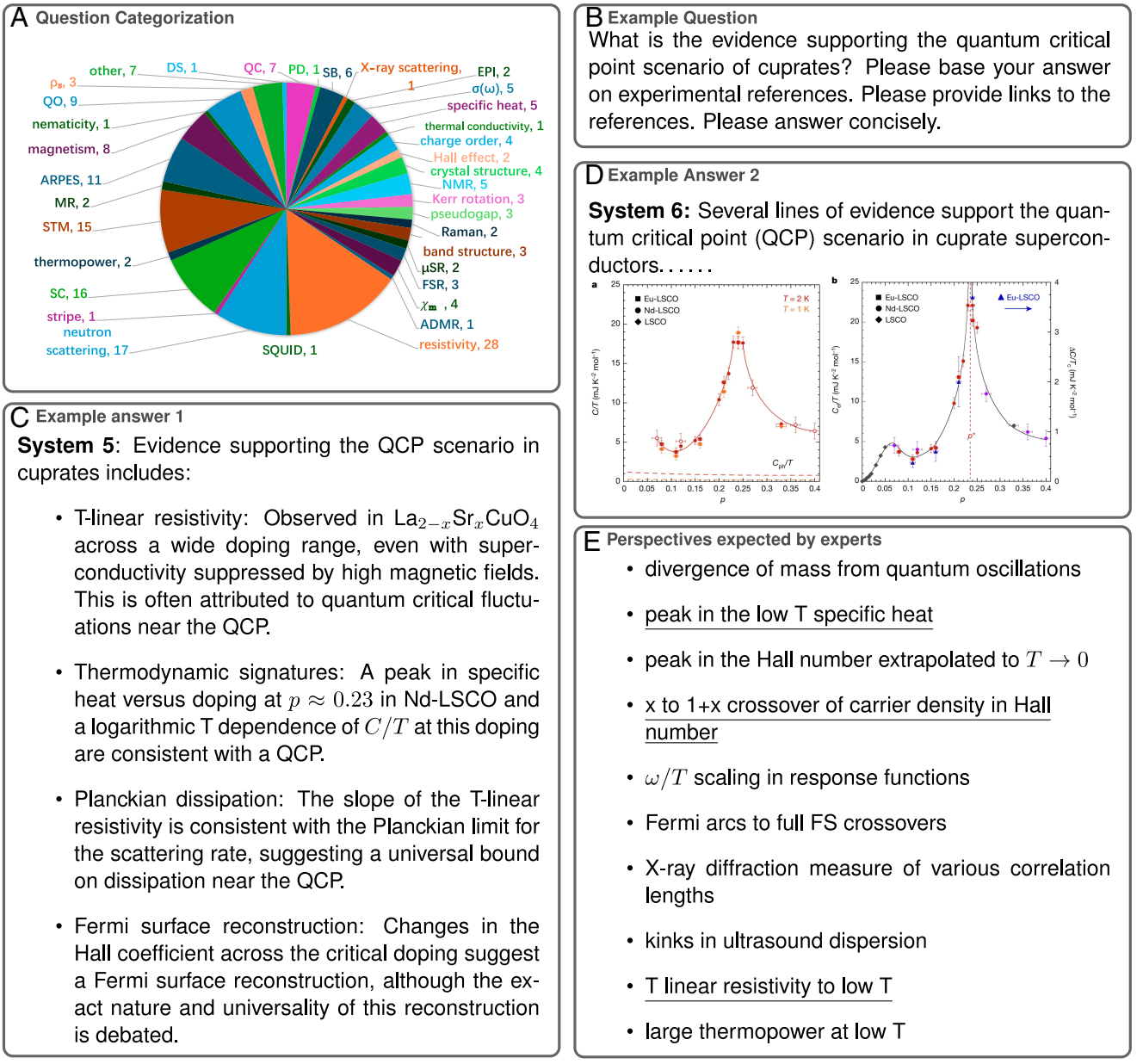


Fig. 2. (*A*) Physical concepts that are involved in the question database and their countings. Each question can be related to multiple concepts. Abbreviations used: ARPES (angle-resolved photoemission), FSR (Fermi surface reconstruction), STM (scanning tunneling microscope), NMR (NMR), MR (magnetoresistance), SC (superconductivity), SQUID (superconducting quantum interference device), ADMR (angle-dependent magnetoresistance), χ_m (magnetic susceptibility), μSR (muon spin rotation/relaxation), $\sigma(\omega)$ (optical conductivity), EPI (electron–phonon interaction), SB (symmetry breaking), PD (penetration depth), QC (quantum criticality), DS (diamagnetic susceptibility), ρ_S (superfluid stiffness), and QO (quantum oscillation). (*B*) A prompt that queries about one question of the database. (*C*) An excerpt of the response to the prompt in (*B*) from System 5 (NotebookLM), which bases its answer on the curated literature database and is instructed to provide multiple perspectives. (*D*) An excerpt of the response to the prompt in (*B*) from System 6 (custom), which bases its answer on the curated literature database and is able to provide figure references. The figures are reprinted from ref. 23 with permission from Springer Nature. The responses in (*C* and *D*) are trimmed for presentation, and the full response is included in *SI Appendix*. (*E*) Perspectives that the expert panel expected to address the question in (*B*). The underlined perspectives are mentioned in the LLM responses.

should survey the relevant experimental literature related to the topic.

Succinctness - Relatively brief and clear answer and explanation of the answer. The response is concise and not rambling or repetitive.

Supported by evidence - The response is based on a collection of experimental evidence reported in the literature. A good response should be based on trustworthy experiments and it should responsibly and comprehensively cite the sources that reported the evidence.

Relevance of images - We could only apply this rubric to two systems capable of surfacing images at the time of the response collection: Perplexity and the Custom system. A good response includes experimental data visualization that supports the claim, should retrieve relevant data visualization of measurement outcomes from the experimental literature and should use them to address the query.

Comments - Observations or comments beyond the above rubric from the expert evaluators.

Downloaded from https://www.pnas.org by Ewha Womans University on September 28, 2026 from IP address 203.255.190.49.

AI Systems for Literature-Based Question Answering

Our goal is to evaluate AI systems' ability to answer these questions accurately. In this study, we included four closed LLM systems that address queries based on training and web search. They are ChatGPT-4o (System-1), Perplexity (System-2), Claude 3.5 (System-3), and Gemini Advanced Pro 1.5 (System-4). We compared the above models with two systems that answer the queries based on our curated literature. The first is NotebookLM (System-5), a Google product that answers users' questions based on 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 Supporting Information (*SI Appendix*, Fig. S4). However, NotebookLM cannot consistently extract figures from 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.

Closed LLM-Based Search Engines. We use 4 popular closed LLM-based methods with web search enabled. These are i) System 1: ChatGPT, ii) System 2: Perplexity, iii) System 3: Claude, and 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.

NotebookLM (System-5). Our fifth system is NotebookLM,* a Google product that answers users' questions based on 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 1,726 papers. 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").

Since NotebookLM is a consumer-oriented product, its responses are targeted toward a lay audience. To get the system to produce language suitable for consumption by scientifically knowledgeable readers, 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 countervailing perspectives in the experimental literature, we 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. An excerpt of NotebookLM response is shown in Fig. 2*C*.

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. *SI Appendix*, Fig. S1 illustrates the full system. The overall system consists of an interface for the user to enter a query and view the responses, as well as a RAG agent that can retrieve relevant documents using an index, and then compose a response based on the extracted information. The final component is an image retriever that uses the query and the retrieved documents to also identify and surface figures that might be relevant.

Building an index. We build a simple multimodal-RAG system using separate text-only and multimodal embedding models for indexing. Relying entirely on multimodal embedding models is insufficient since they do not match the performance of dedicated text-only embedding models on text retrieval tasks (24, 25) as evidenced in document and visual document retrieval benchmarks (26, 27). Accordingly, we use text-only embeddings (28) for indexing and retrieving text chunks, and the open source† multimodal ALIGN embedding model (29) for retrieving and matching images from the source documents associated with the retrieved text and the query. This also makes it much more efficient to retrieve images when the text document source content is quite large since the embedding dimension for text-only is relatively smaller, and does give better text accuracy.

To build the index, we first parsed 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 (30). We then chunked the text and used the text-only embedding model (27) to embed and build an index. We tokenize each full document and then split it into chunks of 4096 tokens, with a 32-token overlap between adjacent chunks. For the images and captions, we used the multimodal embedding model (29) 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 with image retriever. To generate responses for any given query, we first used the index built on the text chunks to retrieve relevant passages from the source papers. The input query is embedded using the same text-only model as the indexed chunks, and the retrieval is based on cosine similarity between the query embedding and the chunk embeddings. We then used the Gemini 1.5 Flash model to compose a coherent response (*SI Appendix*, Fig. S4 shows the prompt) based on the retrieved passages and have the model cite the relevant source papers based on the passages. We then embedded the response and the query using the text-embedder of the multimodal ALIGN model, and took the mean of the query and response texts. We again used cosine similarity to identify the top 5 image feature vectors closest to the combined query-response vector. The final answer from the system consisted of the top-5 retrieved images and the response text along with reference to the source papers (*SI Appendix*, Fig. S2). An example of the response is shown in Fig. 2*D*.

Evaluation of the Responses

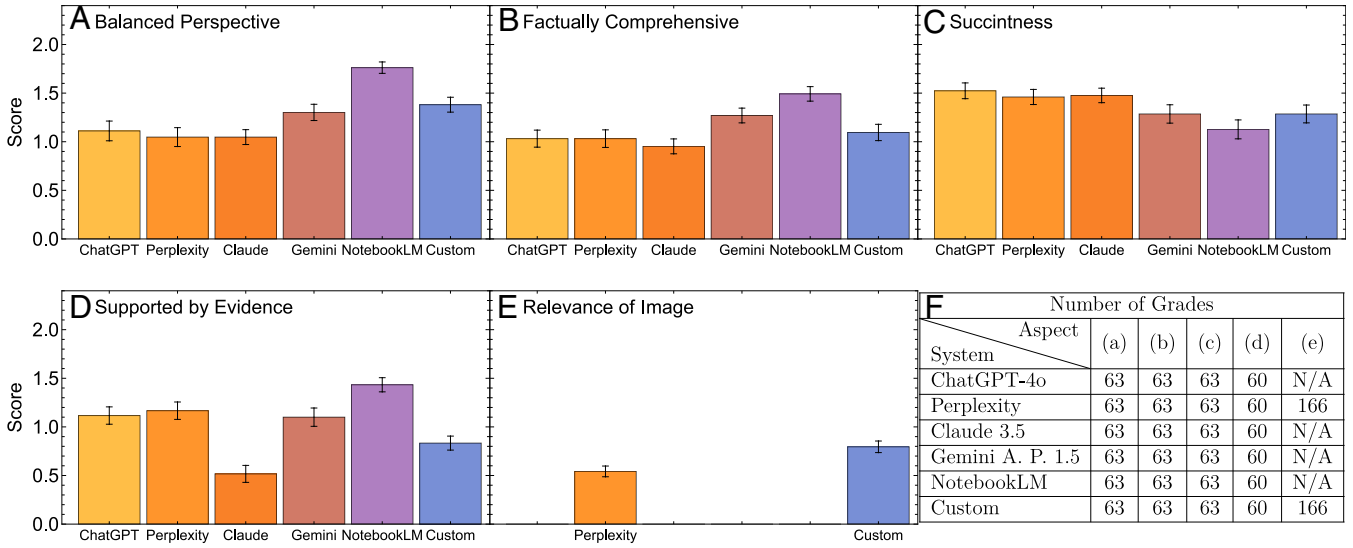
We then evaluated the responses from these six different LLM-based systems on the questions and answers. The responses were collected early December 2024, and compiled on December 12, 2024. We sent the responses back to the experts for evaluation and feedback, with the names of the systems blinded to the experts.

The experts used the questions and the rubric outlined above. Except for the 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

*notebooklm.google.com.

†https://huggingface.co/docs/transformers/model_doc/align.

Downloaded from https://www.pnas.org by Ewha Womans University on September 28, 2026 from IP address 203.255.190.49.



F Number of Grades

System \ Aspect	(a)	(b)	(c)	(d)	(e)
ChatGPT-4o	63	63	63	60	N/A
Perplexity	63	63	63	60	166
Claude 3.5	63	63	63	60	N/A
Gemini A. P. 1.5	63	63	63	60	N/A
NotebookLM	63	63	63	60	N/A
Custom	63	63	63	60	166

Fig. 3. (*A–E*): Mean scores and SEs of the 6 models in 5 aspects: (*A*) Balanced perspective; (*B*) Factually Comprehensive; (*C*) Succinctness; (*D*) Supported by Evidence; (*E*) Relevance of Image. (*F*): The number of grades that enter into the statistics of results in (*A–E*).

experts, with each expert evaluating a subset of the 67 questions. The fifth aspect, the relevance of images, was evaluated by two experts who have reviewed most of the questions. To compare between different models, we only retained scores such that the expert had graded the same (question, aspect) pair across all models. The resulting distribution of expert evaluations is presented in Fig. 3*F*, organized by system and aspect. For each aspect and each system, we calculate the mean and SD of the grades across all questions and experts, as shown in Fig. 3 *A–E*. See *SI Appendix*, Fig. S8 for the distribution of grades on the 0,1,2 scale.

Results. As depicted in Fig. 3 *A*, *B*, and *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*). While our custom system also utilized the curated literature database, it lagged behind NotebookLM in text-based responses, which we attribute to the custom system employing only a simple text retriever. Regarding image retrieval capabilities, only two of our six systems consistently delivered image outputs at the time of the study: Perplexity and our custom system. Between the two systems, the custom system showed superior performance as shown in Fig. 3*E*. Perplexity included schematic sketches or artistic renderings from presentations available on the internet. The custom system (System 6) retrieved figures from the literature data collection to support the response, as illustrated in the example in Fig. 2*D* and *SI Appendix*, Fig. S6. These results above are statistically significant as illustrated in *SI Appendix*, Table S1, 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.

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 (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 nonexperts to the unsettled nature of many topics in the field. An example response is shown in Fig. 2*C*, with more examples in *SI Appendix*, Figs. S5 and S7. Regarding the system's ability to recall and use information from data visualizations, the custom system's responses were superior and more trustworthy because it limited its sources to our literature database (Fig. 2*D* and *SI Appendix*, Fig. S6) for an example prompt and response). However, even the custom system fell short of the expert's needs in this critical capability, as it was unable to quantitatively reason with the data visualization. Both models had to rely on the author's interpretation, as expressed in the text, rather than critically analyzing and absorbing information conveyed through data visualization.

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 and limited perspectives: LLMs often relied on superficial textual similarity rather than conceptual relevance. Even systems which used a curated database, exhibited this issue. For example, it failed to identify key references relevant to quantum criticality, despite those sources being present in the database (Fig. 2*C*). These missed references did not explicitly mention quantum critical points, indicating that the models struggle to recognize implicit conceptual connections. In contrast, human experts understand

Downloaded from https://www.pnas.org by Ewha Womans University on September 28, 2026 from IP address 203.255.190.49.

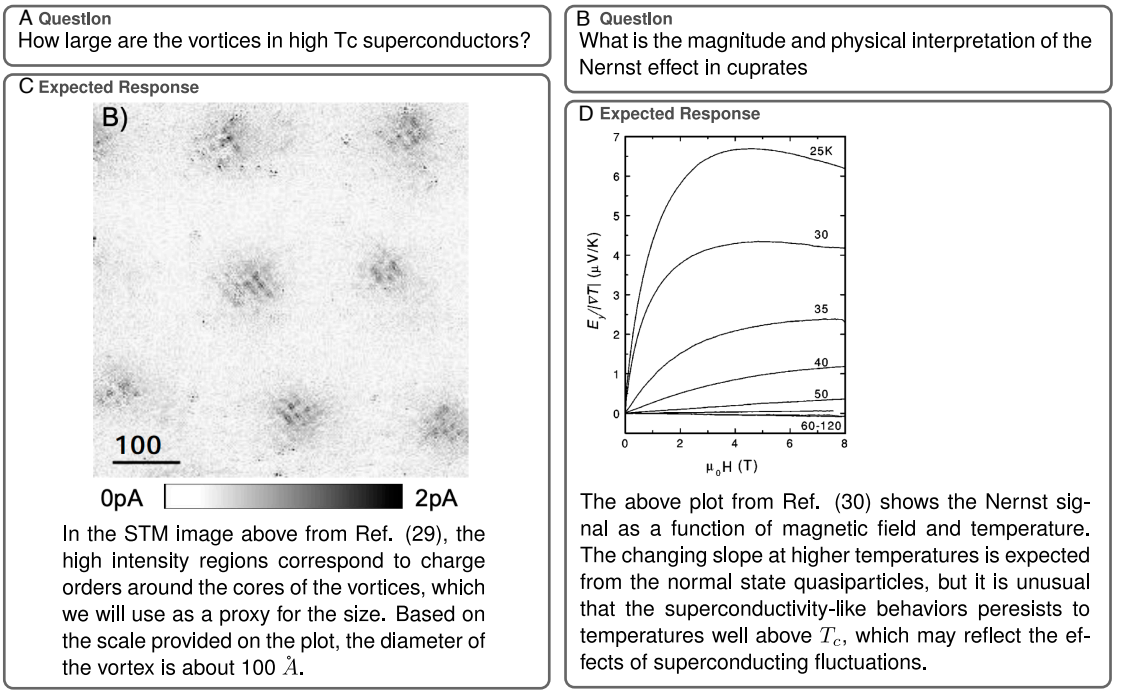


Fig. 4. Two examples showing the expectations of visual reasoning capabilities for future LLMs. (*A* and *B*) are example questions from the question database. (*C* and *D*) are the expected responses, in which the LLMs are expected to surface relevant images and reason based on the contents of the image. The figure in (*B*) is reprinted from ref. 31 with permission from American Association for the Advancement of Science. The figure in (*D*) is reprinted from ref. 32 with permission from the American Physical Society Copyright (2001).

the intrinsic conceptual link between different experiments and desire a more comprehensive survey (Fig. 2*E*).

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 (*SI Appendix*, Fig. S7).

Inaccurate citations: LLMs sometimes supported otherwise reasonable answers with references unrelated to the topic. For example, in *SI Appendix*, Fig. S7, it includes citations to materials not relevant to cuprate superconductors.

Unqualified or biased sources: Systems 1 to 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 nonscientific 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.

Therefore, enhancing the visual reasoning capability is a major direction of improvement for next-generation LLMs. To elaborate on the expectations, we provide two concrete examples below based on questions in our database, which is also shown in Fig. 4.

(i) For the question "How large are the vortices in high-T_c superconductors?," the answer should depend on the experimental technique and the proxy used to extract the length scale. An expert will recognize such unspoken context and answer the question quoting one or more measurements. In Fig. 4, we provide an example figure that can be quoted as a response, adapted from figure 1b of ref. 31. It utilized scanning tunneling microscope (STM) to capture the "halo" of vortices in cuprates, using the local density of states as a proxy. From the intensity plot and the scale bar, it is clear that the low-energy local density of states in the vicinity of vortices extend over surprisingly large length scale of 100 Å. Alternative experimental determinations (e.g., real-space magnetic imaging or independent estimates of coherence length ξ and penetration depth λ) would also be considered acceptable evidence to ground the response. Even when LLMs have access to figures like one shown above, LLMs lacked ability to reason based on the figure to recognize that the scale bar gives quantitative meaning to the image.

(ii) For the question "What is the magnitude and physical interpretation of the Nernst effect in cuprates?," a desired answer would be to offer plots such as figure 5 of ref. 32 and observe that the scale of the Nernst signal for $Bi_2Sr_{2-y}La_yCuO_6$ is at the order of $\mu V/K$ that extends to temperatures as high as $50K$ well above the superconducting transition temperature. Changing slope at higher temperatures is expected from the normal state quasiparticles, but the superconducting state-like behavior at temperatures well above T_c is highly unusual.

Discussions and Outlook

The evolving landscape of AI tools for scientific research encompasses both versatile LLMs and specialized applications. These range from General purpose LLMs (GPT, Claude, Gemini, etc.) that excel in answering simple questions, code generation,

Downloaded from https://www.pnas.org by Ewha Womans University on September 28, 2026 from IP address 203.255.190.49.

and even drafting texts. Any of these are additionally integrated with agentic workflows and web search capabilities, sometimes called "Deep Research," to provide a more in-depth review of topics based on documents, conversations, and resources available on the web. There are also several simple strategies that can be used to improve the custom RAG system. E.g., it would be interesting to improve the retrieval system by including a reranker to rank retrieved text sources on both the image and text chunks. Further, while we use a simple generation prompt that requests the model to surface multiple perspectives, other prompting strategies such as few-shot prompts are worth exploring. Another important aspect to consider is whether we might even want to train or tune the underlying LLM systems, using open source models and strategies such as LoRA (33) so that they can be adapted to the domain and task on a smaller set of text data. These tuning strategies, and agentic workflows will likely require some thought depending on the licensing and availability of the scientific documents.

A major question is whether these systems can be used as specialized research assistants, operating at the level of world experts. Our study is unique in that we have examined this in detail in the context of a specialized and unsolved problem in physics, understanding the origins and mechanisms of high-temperature superconductivity in cuprates. This is a technical field with a finite but rich literature. Among the authors of this paper are some of the leading experts in the field. The experts were able to 1) identify the complete literature in the field, 2) write questions probing deep understanding of the literature, and 3) evaluate the ability of a set of LLM systems to answer them. There are only a small number of people in the world capable of answering questions at the level we are probing, and so this is truly measuring at the level of expert performance.

The results showed that current AI systems fall significantly short on this task. While for foundational or introductory purposes, LLM systems may serve as a useful springboard, they currently lack the ability to distinguish central theoretical frameworks from peripheral ideas. This makes them unsuitable for serious scholarly use without expert oversight. A critical shortcoming is their inability to meaningfully use scientific data visualization in the literature as a reliable source of information on its own right. This severely limits their reliability and utility in answering deep questions about high-temperature superconductivity. An additional challenge is the need to synthesize reported facts across multiple experimental probes, each of which comes with specialized terminology, control knobs, and a dynamic range of operation. LLMs in this study could only surface relevant literature through text matching, but had difficulty identifying conceptual links. Another consistent observation reached by our expert evaluators is that LLMs conflate speculative claims with accepted scientific consensus. This is presumably because they are trained using unvetted internet content, including non-peer-reviewed or fringe material. Finally, the tone of response tends to be authoritative; yet even subtle inaccuracies can mislead nonexperts and obscure the true state of scientific understanding.

A major conclusion of this work is that grounding answers in the experimental literature improved their quality. When models were provided with context of the entire relevant literature, and asked to answer with support from these sources, the quality definitively improved in a blind test. This is reassuring as a conclusion, as it points the way toward more capable expert systems.

One major limitation of this study is how difficult it is to put together this type of evaluation. One needs a finite field and a set of world experts that are able to pose questions and grade answers on topics that correspond to this expertise. Getting bandwidth from such a group is highly nontrivial. Grading responses across our rubric requires expert evaluation, and does not scale either to other fields or even to newer models. To that end, we acknowledge that the evaluation used in this study is not up to date. The evaluations shown in this paper were carried out in early 2025, and so the LLM systems producing the answers were those available at late 2024. Since then, we have seen dramatic progress in the abilities of LLM on the standard rubric of evaluations REFS. Between completing the evaluation and drafting the manuscript, while several expert panel members noted enhancements in text processing performance in newer model iterations, our broader impression is that these models continue to lack the capacity to retrieve data visualizations, underscoring a critical area for future development of LLMs.

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

Note Added. This work is intended as an academic benchmark of current LLM capabilities using a blinded expert evaluation, and is not presented as groundwork for proprietary product development.

Data, Materials, and Software Availability. Some study data are available: The LLM responses of system 5 and system 6 cannot be shared based on legal counsel since there may be possibilities that systems might reproduce content from papers that have copyright restrictions. All other data are included in the manuscript and/or supporting information.

ACKNOWLEDGMENTS. We thank Oliver King and Wesley Hutchins for collaboration and help with NotebookLM. This research is funded in part by the Gordon and Betty Moore Foundation's Emergent Phenomena in Quantum Systems Initiative, Grant GBMF10436 to E.-A.K. H.G. is supported by the Bethe-Wilkins-Kavli Institute at Cornell postdoctoral fellowship of Cornell University and by GBMF10436. E.-A.K. is supported by the NSF through the grant OAC-2118310 and DMR-2433348. J.M.T is supported at Brookhaven by the Office of Basic Energy Sciences, Materials Sciences and Engineering Division, U.S. Department of Energy under Contract No. DE-SC0012704. B.J.R. is supported by NSF Award No. 2428169. The Flatiron Institute is a division of the Simons Foundation.

Author affiliations: [a]Department of Physics, Cornell University, Ithaca, NY 14853; [b]Google, Mountain View, CA 94043; [c]Department of Physics, Harvard University, Cambridge, MA 02138; [d]William H. Miller III Department of Physics and Astronomy, The Johns Hopkins University, Baltimore, MD 21218; [e]Center for Computational Quantum Physics, Flatiron Institute, New York, NY 10010; [f]Collège de France, 75005 Paris, France; [g]The Center for Theoretical Physics (CPHT), CNRS, Ecole Polytechnique, Institut Polytechnique de Paris, Route de Saclay, 91128 Palaiseau, France; [h]Department of Quantum Matter Physics (DQMP), Université de Genève, Genève CH-1211, Switzerland; [i]Université Paris-Saclay, CNRS, Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA), Institut de physique théorique, 91191 Gif-sur-Yvette, France; [j]Department of Physics, Stanford University, Stanford, CA 94305; [k]Physics Program and Initiative for the Theoretical Sciences, The Graduate Center, The City University of New York (CUNY), New York, NY 10016; [l]Department of Physics and Astronomy, College of Staten Island, The City University of New York (CUNY), Staten Island, NY 10314; [m]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139; [n]Condensed Matter Physics and Materials Science Division, Brookhaven National Laboratory, Upton, NY 11973-5000; [o]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138; and [p]Department of Physics, Ewha Womans University, Seoul 03760, South Korea

Author contributions: H.G., M.A., E.D., M.P.B., S.V., and E.-A.K. designed research; H.G., M.T., P.R., A.V., C.C., D.J.L., S.E., M.A., E.D., N.P.A., C.F., A.G., O.G., D.K., S.A.K., V.O., B.J.R., S.S., T.S., J.M.T., M.P.B., S.V., and E.-A.K. performed research; H.G., M.A., E.D., N.P.A., C.F., A.G., O.G., D.K., S.A.K., V.O., B.J.R., S.S., T.S., J.M.T., M.P.B., S.V., and E.-A.K. analyzed data; and H.G., M.A., E.D., N.P.A., O.G., S.A.K., B.J.R., S.S., J.M.T., M.P.B., S.V., and E.-A.K. wrote the paper.

Reviewers: M.B., University of Maryland, College Park; M.-H.J., Laboratoire National des Champs Magnetiques Intenses–Grenoble; and Y.-Z.Y., University of California, San Diego.

Downloaded from https://www.pnas.org by Ewha Womans University on September 28, 2026 from IP address 203.255.190.49.

1. J. G. Bednorz, K. A. Müller, Possible highTc superconductivity in the Ba-La-Cu-O system. *Z. Phys. B Condens. Matter* **64**, 189–193 (1986).
2. M. K. Wu *et al*., Superconductivity at 93 K in a new mixed-phase Y-Ba-Cu-O compound system at ambient pressure. *Phys. Rev. Lett.* **58**, 908–910 (1987).
3. H. Maeda, Y. Tanaka, M. Fukutomi, T. Asano, A new high-Tc oxide superconductor without a rare earth element. *Jpn. J. Appl. Phys.* **27**, L209 (1988).
4. Z. Z. Sheng, A. M. Hermann, Superconductivity in the rare-earth-free Tl–Ba–Cu–O system above liquid-nitrogen temperature. *Nature* **332**, 55–58 (1988).
5. S. N. Putilin, E. V. Antipov, O. Chmaissem, M. Marezio, Superconductivity at 94 K in $HgBa_2Cu0_{4+\delta}$ *Nature* **362**, 226–228 (1993).
6. H. Takagi, S. Uchida, Y. Tokura, Superconductivity produced by electron doping in CuO_2-layered compounds. *Phys. Rev. Lett.* **62**, 1197–1200 (1989).
7. C. M. Varma, Colloquium: Linear in temperature resistivity and associated mysteries including high temperature superconductivity. *Rev. Mod. Phys.* **92**, 031001 (2020).
8. D. F. Agterberg *et al*., The physics of pair-density waves: Cuprate superconductors and beyond. *Ann. Rev. Condens. Matter Phys.* **11**, 231–270 (2020).
9. C. Proust, L. Taillefer, The remarkable underlying ground states of cuprate superconductors. *Ann. Rev. Condens. Matter Phys.* **10**, 409–429 (2019).
10. E. Fradkin, S. A. Kivelson, J. M. Tranquada, Colloquium: Theory of intertwined orders in high temperature superconductors. *Rev. Mod. Phys.* **87**, 457–482 (2015).
11. S. E. Sebastian, C. Proust, Quantum oscillations in hole-doped cuprates. *Ann. Rev. Condens. Matter Phys.* **6**, 411–430 (2015).
12. N. P. Armitage, P. Fournier, R. L. Greene, Progress and perspectives on electron-doped cuprates. *Rev. Mod. Phys.* **82**, 2421–2487 (2010).
13. L. Taillefer, Scattering and pairing in cuprate superconductors. *Ann. Rev. Condens. Matter Phys.* **1**, 51–70 (2010).
14. T. P. Devereaux, R. Hackl, Inelastic light scattering from correlated electrons. *Rev. Mod. Phys.* **79**, 175–233 (2007).
15. P. A. Lee, N. Nagaosa, X. G. Wen, Doping a Mott insulator: Physics of high-temperature superconductivity. *Rev. Mod. Phys.* **78**, 17–85 (2006).
16. D. N. Basov, T. Timusk, Electrodynamics of high-T_c superconductors. *Rev. Mod. Phys.* **77**, 721–779 (2005).
17. G. Deutscher, Andreev-Saint-James reflections: A probe of cuprate superconductors. *Rev. Mod. Phys.* **77**, 109–135 (2005).
18. S. A. Kivelson *et al*., How to detect fluctuating stripes in the high-temperature superconductors. *Rev. Mod. Phys.* **75**, 1201–1241 (2003).
19. S. Sachdev, Colloquium: Order and quantum phase transitions in the cuprate superconductors. *Rev. Mod. Phys.* **75**, 913–932 (2003).
20. A. Damascelli, Z. Hussain, Z. X. Shen, Angle-resolved photoemission studies of the cuprate superconductors. *Rev. Mod. Phys.* **75**, 473–541 (2003).
21. C. C. Tsuei, J. R. Kirtley, Pairing symmetry in cuprate superconductors. *Rev. Mod. Phys.* **72**, 969–1016 (2000).
22. S. Pramanick, R. Chellappa, S. Venugopalan, SPIQA: A dataset for multimodal question answering on scientific papers. *NeurIPS* **37**, 118807–118833 (2024)
23. B. Michon *et al*., Thermodynamic signatures of quantum criticality in cuprate superconductors. *Nature* **567**, 218–222 (2019).
24. R. A. Benedikt Schifferer, G. Moreira, Best-in-class multimodal rag: How the llama 3.2 nemo retriever embedding model boosts pipeline accuracy (2025). https://developer.nvidia.com/blog/best-in-class-multimodal-rag-how-the-llama-3-2-nemo-retriever-embedding-model-boosts-pipeline-accuracy. Accessed 23 February 2026.
25. M. Faysse *et al*., Colpali: Efficient document retrieval with vision language models. arXiv [Preprint] (2024). https://arxiv.org/abs/2407.01449 (Accessed 23 February 2026).
26. VTI Technology, Vidore leaderboard: Visual document retrieval evaluation (2024). https://huggingface.co/spaces/vidore/vidore-leaderboard.
27. Q. Macé, A. Loison, M. Faysse, Vidore benchmark v2: Raising the bar for visual retrieval. [Preprint] (2025). https://arxiv.org/abs/2505.17166 (Accessed 23 February 2026).
28. J. Lee *et al*., Gecko: Versatile text embeddings distilled from large language models. arXiv [Preprint] (2024). http://arxiv.org/abs/2403.20327. Accessed 23 February 2026.
29. C. Jia *et al*., "Scaling up visual and vision-language representation learning with noisy text supervision" in *International Conference on Machine Learning*, M. Meila, T. Zhang, Eds. (PMLR, 2021), pp. 4904–4916.
30. C. Clark, S. 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*, N. R. Adam, L. Cassel, Y. Yesha, R. Furuta, M. C. Weigle, Eds. (Association for Computing Machinery, New York, NY, USA, 2016), pp. 143–152.
31. J. E. Hoffman *et al*., A four unit cell periodic pattern of quasi-particle states surrounding vortex cores in $Bi_2Sr_2CaCu_2O_{8+\delta}$. *Science* **295**, 466–469 (2002).
32. Y. Wang *et al*., Onset of the vortexlike nernst signal above T_c in $La_{2-x}Sr_xCuO_4$ and $Bi_2Sr_{2-y}La_yCuO_6$. *Phys. Rev. B* **64**, 224519 (2001).
33. E. J. Hu *et al*., Lora: Low-rank adaptation of large language models. *ICLR* **1**, 3 (2022).

Downloaded from https://www.pnas.org by Ewha Womans University on September 28, 2026 from IP address 203.255.190.49.