Project Name: MatSci-GPT: *The go-to chat-bot for materials scientists*
Team Number: 51
Team Members: Milan Jain, Kendrick Hsu, Jon Wong, Peter Li, Jeremy Zein
Faculty Advisors: Delip Rao, Chris Callison-Burch

Executive Summary:

Team 51 is a group of highly motivated students with a passion for deploying AI and large language model (LLM) technology to tackle problems in the Educational Technology (EdTech) sector. We were brought together by this joint interest and intend to leverage our diverse set of technical expertise to solve a challenge that we believe affects materials science students across the country.

We have come together to build MatSci-GPT: an interactive, AI-powered chatbot built to be fit for the materials science community, which we define to include students, researchers, and professionals in academia and industry. Relative to existing general-purpose AI tools that many use today, MatSci-GPT is trained on a body of over 5,000 materials science research papers, with new publications added to the database on a bi-monthly basis via an automated pipeline that we built in-house. The platform incorporates Retrieval-Augmented Generation (RAG) to provide accurate, source-cited responses - this enables users to interact with scientifically reliable knowledge.

In addition, we believe that MatSci-GPT differentiates itself relative to competitors via its domain-specific image generation capabilities for visualizing unit cells as well as phonon band structures, which are fundamental components of any materials science analysis. Based on our own experiences, these features are either unavailable or inaccurate in the majority of chatbot platforms that exist on the market today.

The product addresses three shortcomings left by existing technology: first, a lack of domain-specific expertise; second, an absence of verifiable source attribution; and third, an inability to generate accurate scientific images. The product delivers a strong and consistent value proposition to a variety of stakeholders - these include materials science students looking for support, faculty researchers, and industry professionals.

Driven by resilient and concisely growing demand in both the EdTech and R&D sectors, MatSci-GPT is well positioned to serve a sizable TAM market. With by a scalable revenue model - including freemium plans, licenses, and academic partnerships - MatSci-GPT in its full form can become the go-to research tool in the field of materials science. Its combination of source retrieval, hyper specific functionality, and scientific accuracy offers a compelling value proposition its end users.

Value Proposition:

MatSci-GPT is an interactive, AI-powered chatbot build for the materials science community - including students, researchers, and practitioners. Unlike general-purpose LLMs, MatSci-GPT is trained on a specialized set of roughly 5,000 up-to-date domain relevant research papers and automatically updates the database to include the newest publications via a bi-monthly scraping and preprocessing pipeline. Leveraging RAG, it generates accurate, source-cited responses

based on the relevant peer-reviewed scientific literature stored in the database. This process ensures a combination of transparency and academic rigor.

The platform addresses three gaps in existing chatbot solutions available in the market: (**1) lack of domain-specific expertise, (2) absence of reliable source attribution**, and (3) **inability to handle** materials science-specific tasks such as **image generation** of crystal unit cells and phonon band structures.

First and foremost, general-purpose chatbots like ChatGPT are trained on large and heterogeneous datasets intended to maximize versatility across many fields. However, this breadth comes at the cost of depth, in particular in the contexts of highly specialized scientific domains like materials science. These models often fail to grasp nuanced terminology, emerging concepts, and interdisciplinary connections unique to the field. For example, they may confuse similar-sounding materials, misinterpret notation in phase diagrams, or give outdated explanations of certain techniques. For students and researchers, these inaccuracies are significant as they undermine understanding and can, in the worst case, lead to the propagation of misinformation. MatSci-GPT addresses this issue by training exclusively on materials science literature, ensuring a knowledge base that reflects both foundational concepts and the latest research developments.

Second, one of the most significant shortcomings of existing chatbot solutions is their inability to provide source attribution for their answers. When a model responds without citing where its information originated, end users are left uncertain about the response's credibility. One can see how this would be a major issue in academic and research contexts. The ambiguity makes it difficult to trust and therefore use or build upon chatbot outputs, particularly in the field of science where precision is a must. MatSci-GPT overcomes this through RAG, a system that retrieves and references peer-reviewed sources from its specifically curated database, which improves trust in the system.

Third and finally, another restriction of traditional LLM technology is its inability to provide reliable image generation, an important asset in technical disciplines. In the context of materials science, visual representations such as unit cells and phonon band structures are important for highlighting structural relationships. Most chatbots either can't generate accurate images or fail to support integrations with domain-specific visualization tools. The result of this is forcing users to switch platforms or search manually with a traditional Google search. MatSci-GPT addresses this by allowing users access to AI-based image generation that has been adapted to the materials science domain. In this way, our product lets users generate scientifically accurate images on demand.

Individuals in the materials science community have tried to work around the limitations of general purpose chatbots. For example, many materials science students have attempted to adapt these tools by creating custom GPTs. This process involves uploading specific material such as journal articles, lecture notes, and research papers relevant to the area of study at hand. In our view, this approach can in certain cases provide improvements in relevance, but it

largely remains an onerous and insufficient process. Collecting, formatting, and uploading high-quality domain-specific materials takes time and effort. Second, the quality and accuracy of the chatbot's responses often remain low, especially when interpreting complex data types like equations, crystal structures, or scientific notation. Third, most of these solutions still lack sufficient retrieval capabilities, making it difficult for users to verify the sources of responses or trace back to original references.

MatSci-GPT directly addresses these pain points by automating domain-specific knowledge ingestion and enabling precise, source-cited responses via its real-time RAG system. Rather than requiring users to upload materials, MatSci-GPT does that work by scraping and processing the latest research papers in materials science. The platform is designed and tailored to the needs of the materials scientist, offering capabilities such as crystal structure visualization, band structure image generation. By combining text-based reasoning with visual generation and reliable source tracking, MatSci-GPT aims to be the all-in-one research tool for the materials science community.

Stakeholders:

For our product, we identify three primary stakeholder groups: students and early-career researchers, academic and research faculty, and industry practitioners. We also believe there are some important "peripheral" (secondary) stakeholders: publishers and scientific content providers, and scientific societies and professional organizations.

Students and early-career researchers are the primary users of MatSci-GPT. They rely on our platform for fast, reliable access to cutting-edge scientific information, source-cited responses, and visual aids like unit cell and phonon band structure generation. These users benefit from reduced time spent manually searching literature or creating scientific diagrams, allowing them to focus more on learning, research, and innovation. As our core users, their consistent feedback will be important in refining the user interface, improving feature sets, and ensuring the product remains up-to-date and tailored to the evolving academic and research needs.

Academic and research faculty serve dual roles: endorsers and validators. Faculty members can integrate MatSci-GPT into their teaching or research workflows, offering it as a supplementary tool for students or leveraging it themselves for literature exploration and visualization tasks. Their endorsements will build credibility for the platform, while their feedback on scientific accuracy and domain relevance will be critical to maintaining high-quality outputs that meet rigorous academic standards.

Industry practitioners, including materials engineers, R&D scientists, and technical consultants, represent another important stakeholder group. In industry settings where quick access to reliable scientific information is vital for material selection, process optimization, or innovation efforts, MatSci-GPT provides immediate, trusted assistance. Its ability to generate

accurate visualizations and reference the latest peer-reviewed research can support faster decision-making and reduce reliance on slow manual searches.

With respect to the peripheral stakeholders, **publishers and scientific content providers** play an important role by licensing academic papers, datasets, and imagery that train and enhance the chatbot's capabilities. Their cooperation ensures the chatbot's database remains accurate and up-to-date. Meanwhile, **scientific societies and professional organizations** are potential partners for promoting MatSci-GPT, validating its value, and possibly providing access to specialized content or member engagement channels.

Market Overview and Competition:

MatSci-GPT sits at the intersection of two rapidly growing markets: EdTech and Scientific Research Tools. Its core audience - students, researchers, and professionals in materials science - represents a specialized but globally significant segment within these broader industries. The demand for intelligent, domain-specific tools is accelerating as the volume of scientific publications grows and as academic and industrial users seek efficient ways to consume and visualize complex technical information. The global EdTech market is projected to surpass $400 billion by 2025. This growth is driven by increased adoption of AI-powered tools and remote learning platforms. Within this broader space, STEM education in particular is a key area of investment. MatSci-GPT addresses a major gap in this market: the lack of specialized, research-grade AI tools tailored for deep scientific domains like materials science. At the same time, the scientific research and development (R&D) tools market, which includes software for simulation, data analysis, and literature discovery, is also expanding. Researchers today face significant inefficiencies in literature review, data visualization, and keeping up with the latest findings. With over 100,000 materials science papers published annually, manual scraping is no longer a sustainable option for researchers trying to innovate quickly. MatSci-GPT steps in to meet and solve this need by offering a centralized platform for source-cited answers, domain-specific reasoning, and visual generation capabilities. Furthermore, the rise of LLMs and RAG technology has created a new category of AI assistants with the potential to transform research workflows. MatSci-GPT stands out in this emerging space by combining fine-tuned scientific knowledge, real-time source attribution, and image generation specific to materials science.

The competitive landscape for MatSci-GPT includes general-purpose AI tools like ChatGPT and Perplexity AI, as well as traditional research platforms such as Google Scholar and domain-specific databases like the Materials Project. While general AI platforms offer solid conversational abilities, they lack deep materials science expertise, rigorous source attribution, and the ability to generate domain-specific scientific visuals, making them insufficient for our target user base. Traditional research databases provide access to papers but require users to manually sift through information without real-time summarization, cross-referencing, or visualization support. Emerging custom GPT offerings also exist, but they typically demand significant user effort to upload materials and often struggle with complex scientific concepts.

MatSci-GPT differentiates itself by combining RAG trained on a dynamic, curated body of materials science literature with scientific image generation capabilities for unit cells, phonon band structures, and more. This specialization allows MatSci-GPT to bridge the gap between static research databases and generic AI chatbots, offering a fully integrated solution that is effectively tailored to the needs of the materials science community.

MatSci-GPT faces a few risks that we as owners of the product must proactively manage. One challenge is ensuring the accuracy and reliability of responses in a highly technical domain. As we've noted throughout this report, errors in scientific reasoning, misinterpretation of source material, or incorrect visual outputs could undermine trust among users who need accurate answers. Second,risk of data quality and scope is important: while the RAG system is designed to retrieve from curated scientific papers, maintaining a comprehensive, regularly updated, and high-integrity database will require ongoing infrastructure, which can be costly and time-intensive on our end. User adoption is another persistent risk. The process of convincing researchers, students, and institutions to incorporate a new tool into established workflows requires a clear demonstration of value and usability. Finally, compliance with data privacy regulations - particularly if user accounts or institutional integrations are pursued - will be essential to build institutional partnerships and protect sensitive information.

Customer Segment:

MatSci-GPT is designed for a focused yet diverse customer segment composed of materials science students, academic researchers, and industry professionals. At the student level, this includes both undergraduates and graduate students enrolled in specialized coursework who need support navigating complex, interdisciplinary concepts. These users - based on our primary research - are tech-savvy, highly motivated, and often struggle with accessing up-to-date resources, generating clear visualizations, or understanding how concepts interrelate across physics, chemistry, and biology. First-hand feedback from Penn materials science students underscores this need. One student noted, "ChatGPT is relatively useless for many materials science problems," emphasizing that a model tailored to the discipline's terminology and fundamentals would be significantly more useful. Another student shared that "materials science is sometimes very niche and without reading specific scientific papers it is hard to find information," highlighting the challenge of accessing reliable, contextual information in this field.

In academic research settings, MatSci-GPT supports postdocs and faculty who are pressed for time yet need to stay current with emerging literature and visualize new structures and properties rapidly. These users value the tool's speed, source-cited accuracy, and ability to offload repetitive tasks like literature review or structure rendering. Industry professionals - such as materials engineers, computational scientists, and R&D staff - also represent a valuable segment. These users benefit from MatSci-GPT's ability to quickly gather relevant literature, generate technical visuals, and accelerate discovery cycles in applied contexts. Across all segments, the common value driver is efficiency without compromising accuracy.

Intellectual Property:

MatSci-GPT's core intellectual property lies in its domain-specific fine-tuning and proprietary data processing pipeline. The model's ability to retrieve and cite from a curated, frequently updated database of materials science research papers involves a custom-built pipeline that automates ingestion, cleaning, and indexing of scientific literature - an asset that may be eligible for trade secret protection. The chatbot's visual generation features, including the ability to create unit cell and phonon band structure images from user prompts, may also involve custom image-rendering algorithms or backend architecture that could be protected as utility patents or copyrighted software. In terms of licensing, MatSci-GPT will need to ensure proper use rights for any external content, such as scientific figures, datasets, or third-party APIs used in visualizations. To maintain a competitive edge and ensure legal compliance, a combination of open-source license management, institutional content agreements, and strategic IP filings will be crucial as the product matures.

Cost:

The development of MatSci-GPT involves costs including infrastructure, model training, content acquisition, and product development. At a high level, the most significant recurring costs are associated with cloud computing resources, particularly for hosting the chatbot, maintaining a scalable RAG system, and running inference for image generation models. Building and maintaining a scientific literature database - including automated scraping, parsing, and indexing of thousands of materials science papers - requires robust backend systems and incurs ongoing storage and processing costs. Additionally, model fine-tuning and experimentation on domain-specific data introduces compute-intensive training expenses, especially when using GPU-enabled virtual machines. On the front-end side, developing a secure, user-friendly interface with features like persistent chat histories, user authentication, and preference customization adds design and engineering workload. Other key costs include software licenses, data cleaning tools, and potential licensing fees for access to proprietary papers or visualization libraries.

The abovementioned costs break down approximately as follows:

- Cloud infrastructure (compute, storage, inference APIs): $3,000-5,000/year
- Model training and fine-tuning (GPU compute credits): $2,000-4,000/project phase
- Frontend & backend development tools (software libraries, dev platforms): $500-1,000
- Data acquisition & licensing (scientific sources, content APIs): $1,000-3,000
- Legal and compliance costs (IP protection, FERPA/GDPR considerations): $2,000-5,000
- Miscellaneous (surveys, user testing incentives, minor services): $500-1,000

Based on this, in total, initial development and deployment of MatSci-GPT could be achieved with a budget of **$10,000-20,000**, with additional recurring costs if the platform is scaled institutionally or offered commercially.

Revenue Model:

MatSci-GPT's core revenue strategy will be based on a tiered subscription model, supplemented by enterprise licensing, academic partnerships, and data-access agreements.

For individual users, MatSci-GPT will offer a freemium model. The basic tier would provide access to general chatbot functionality and limited image generation, while premium tiers ($10 - $25/month) would unlock features like unlimited queries, advanced visualization tools (unit cells, phonon band structures), citation export tools, and more. A discounted academic plan for university students could encourage early adoption, with pricing subsidized or bundled through institutional licenses.

For universities, research labs, and corporate R&D departments, MatSci-GPT can offer enterprise-level access via annual contracts. These packages would include administrative dashboards, multi-user support, LMS integration, and API access for embedding chatbot functionality within internal research tools. Pricing for institutional licenses could range from $5,000 to $50,000 per year, depending on user count and feature scope.

In parallel, MatSci-GPT can partner with publishers and academic content providers, offering visibility and engagement analytics in exchange for content access or entering revenue-sharing agreements for licensed materials. Additionally, anonymized usage data (compliant with privacy laws) could be leveraged to generate insight reports for institutions or societies interested in tracking topic trends and user needs in the field of materials science.