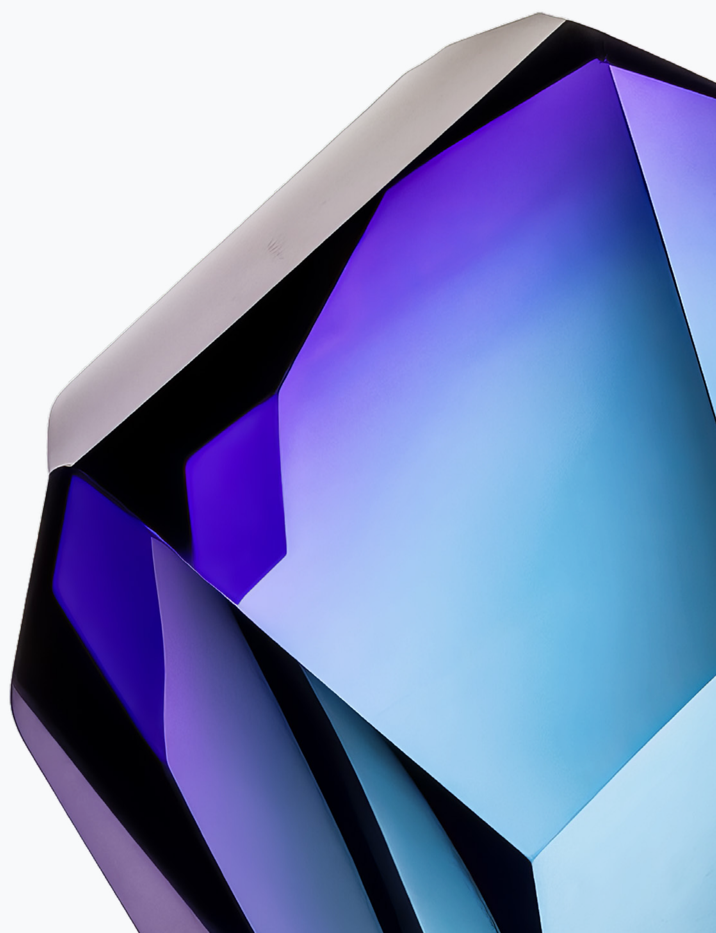
# Solving Domain-Specific Problems Using LLMs

Authors: Christopher Semturs,
Shekoofeh Azizi, Scott Coull,
Umesh Shankar
and Wieland Holfelder



## Acknowledgements

### Reviewers and Contributors

Johan Schalkwyk

Sami Lachgar

Dale Webster

Yun Liu

Connie Fan

Chris Grier

Antonio Gulli

Anant Nawalgaria

### Curators and Editors

Antonio Gulli

Anant Nawalgaria

Grace Mollison

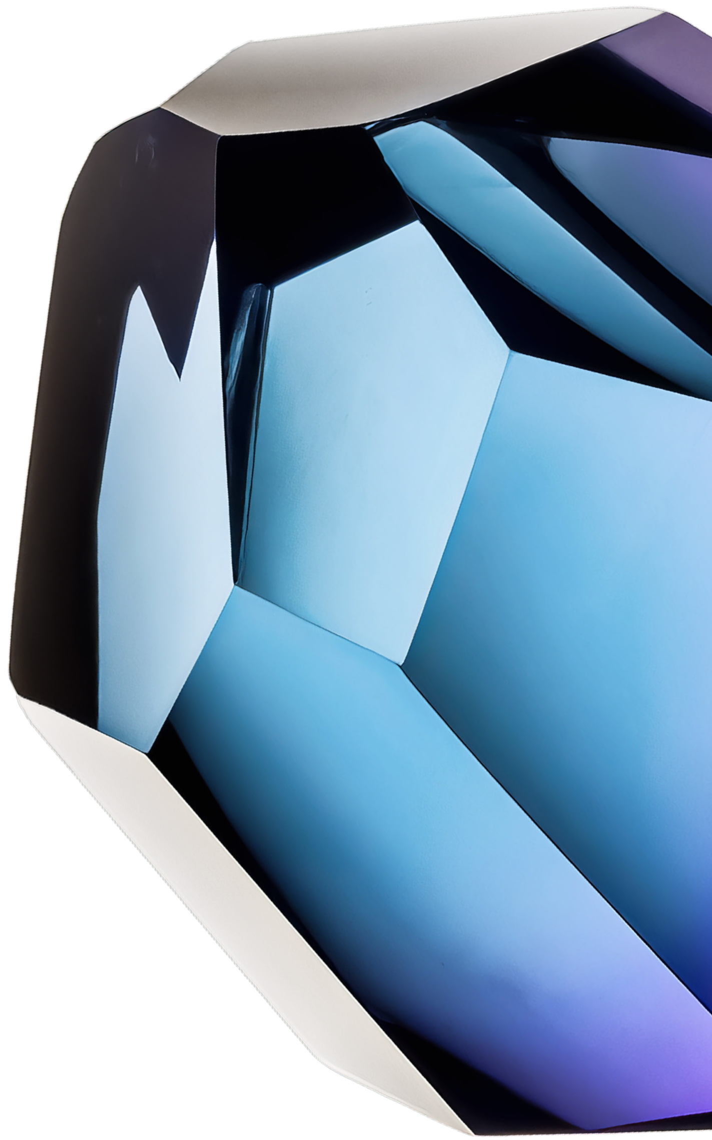
### Technical Writer

Karen Lin

### Designers

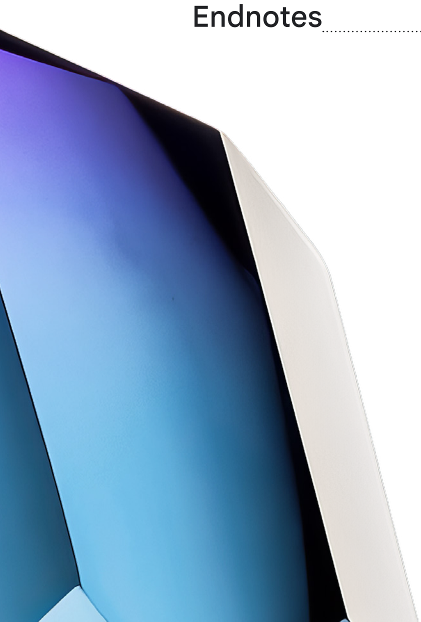
Michael Lanning

Luxi Chen



# Table of contents





> ... recent developments have highlighted the potential of fine-tuning LLMs to address specific problems within specialized fields.

# Introduction

Large language models (LLMs) have emerged as powerful tools for tackling complex challenges in numerous domains. While early iterations focused on general-purpose tasks, recent developments have highlighted the potential of fine-tuning LLMs to address specific problems within specialized fields. This whitepaper explores these concepts in two distinct domains: cybersecurity and medicine. Each showcases the unique ability of LLMs to enhance existing workflows and unlock new possibilities.

Cybersecurity presents a number of unique challenges for LLMs, including a scarcity of publicly available data, a wide diversity of highly technical concepts, and information about threats that change on a daily basis. Additionally, sensitive use cases, like malware analysis, necessitate specific considerations for model development. We address these challenges by focusing on cybersecurity-specific content and tasks, pairing security-focused language models with a suite of supporting techniques to offer improved performance for vital tasks like threat identification and risk analysis.

In the field of medicine, LLMs face a different set of obstacles, such as the vast and ever-evolving nature of medical knowledge and the need to apply said knowledge in a context-dependent manner that makes accurate diagnosis and treatment a continual challenge. LLMs like Med-PaLM, customized for medical applications, demonstrate the ability to answer complex medical questions and provide insightful interpretations of medical data, showing potential for supporting both clinicians and patients.

Through the lens of these two distinct domains, in this whitepaper we will explore the challenges and opportunities presented by specialized data, technical language, and sensitive use cases. By examining the unique paths taken by SecLM and Med-PaLM, we provide insights into the potential of LLMs to revolutionize various areas of expertise.

# SecLM and the future of cybersecurity

Security practitioners face a myriad of challenges, including new and evolving threats, operational toil, and a talent shortage. Specialized Generative AI (Gen AI) can help address these challenges by automating repetitive tasks, freeing up time for more strategic activities, and providing new opportunities to access knowledge.

## Challenges in cybersecurity

In the movies, we often see information security reduced to the caricature of hoodie-clad and headset-wearing hackers with ill intent, armed with ruggedized laptops, tapping away furiously until we hear the two magic words: “I’m in.”

To the extent that you even see the defenders, they are in reactive mode-think war rooms, empty coffee cups, people barking orders, and monitors showing the attacker’s every move in real-time.

That is Hollywood; we live in the real world.

In reality, the people who practice cybersecurity - the developers, system administrators, SREs, and many junior analysts to whom our work here is dedicated - have the Sisyphean task of keeping up with the latest threats and trying to protect complex systems against them. Many practitioners’ days are largely filled with repetitive or manual tasks, such as individually triaging hundreds of alerts, that take valuable time away from developing more strategic defenses. The momentum is definitely not in the defender’s favor; attackers are adopting advanced technologies, including artificial intelligence,[1] to extend their reach and quicken the pace of exploitation. And there are definitely no monitors showing the attacker’s every move!

Based on our experience working with users and partners, we see three major challenges in the security industry today: **threats**, **toil**, and **talent**.

- **New and evolving threats**: The threat landscape is constantly changing, with new and increasingly sophisticated attacks emerging all the time. This makes it difficult for defenders to keep up with the latest information, and conversely for practitioners to sift through that flood of data to identify what’s relevant to them and take action.

- **Operational toil**: People working in security operations or DevOps roles often spend a significant amount of time on repetitive manual tasks that could be automated or assisted. This leads to overload and takes away time from more strategic activities. Excessive focus on minutiae also prevents analysts and engineers from seeing the bigger picture that is key to securing their organizations.
- **Talent shortage**: There is a shortage of skilled security professionals, making it difficult for organizations to find the people they need to protect their data and systems. Often, people enter security-focused roles without much training and with little spare time to expand their skills on the job.

Without the ability to address these three challenges, it will be difficult to keep up with the demands of modern cybersecurity systems.

## How GenAI can tackle the challenges in cybersecurity

We envision a world where novices and security experts alike are paired with AI expertise to free themselves from repetition and toil, accomplish tasks that seem impossible to us today, and provide new opportunities to share knowledge. Large language models (LLMs) and adjacent GenAI techniques can meaningfully improve the working lives of both security novices and experienced practitioners. Indeed, in many cases, we have already found that GenAI is useful to solve a number of real-world security problems in our challenge areas:

| Persona(e) | Challenges faced | How Gen AI can help |
| --- | --- | --- |
| **Security analyst** | Analysts not familiar with each tool's bespoke schema and query language. | Translate natural-language queries into a domain-specific security event query language and rules language. |
| | Investigating, clustering, and triaging incoming alerts is time-consuming and requires multiple steps and tools. | Autonomous capabilities to perform investigation, grouping, and classification, incorporating context and real-time tool use. |
| | Hard to assemble the right series of tailored steps to remediate an issue. | Personalized, case-specific remediation planning in user environments. |
| **Threat Researcher or System Administrator** | An unknown and obfuscated artifact (such as a script or binary) is discovered and can't be easily analyzed manually. | Automated reverse engineering with LLM-powered code analysis with tool use for de-obfuscation and decompilation. Explain, analyze, and classify potentially malicious artifacts. |
| **CISO team** | Manual work required to identify and summarize the most likely threats facing the organization. | Generate a readable document or slide deck, applying the latest threat intelligence and findings from security tools to the specific organization. |
| **IT Administrator Dedicated Security Team** | Hard to understand all the ways an attacker could access sensitive resources. | Identify potential or actual attack paths, highlighting key elements and remediations. |

| | | |
|---|---|---|
| **Application Developers** | Challenging to determine the right places to fuzz-test an application. | Identify which locations to fuzz-test and generate the appropriate code. |
| **Application Developers & IT Administrators** | Keep access policies aligned to the principle of least privilege. | Given historical access patterns and current configuration, construct a configuration file modification that grants a more minimal set of roles. |
| **A person responsible for an application or system** | People don't always understand security concepts or how to apply them to their environments; they have to know how to break a problem down, ask questions in many places, and then combine them to obtain an answer. | Give an answer that reflects authoritative security expertise and, using integrations, is relevant to the user's working environment. |

To tackle these problems in a meaningful and holistic way, however, we need a multi-layered approach:

- **Top layer:** existing security tools that understand the relevant context and data, and can actuate necessary changes;
- **Middle layer:** a security-specialized model API with advanced reasoning and planning capabilities;
- **Bottom layer:** datastores of authoritative security intelligence and operational expertise

Notably, one of the key benefits of LLMs is their ability to process and synthesize vast amounts of heterogenous data – an important capability in the increasingly siloed world of cybersecurity data. We seek to leverage that capability to solve challenging security problems, whether by assisting human analysts or through autonomous agents, by combining relevant context and authoritative sources with a flexible planning framework in a single API, which we call SecLM.

This API offers rich planning capabilities that combine LLMs and other ML models, Retrieval-Augmented Generation (RAG) to ground results in authoritative data, and tool use to perform actions or look up relevant information. We argue that this holistic approach is critical because accuracy is so important in security and LLMs alone cannot inherently solve all security problems.

## SecLM: An API for cybersecurity tasks

Our vision of the SecLM API is to provide a 'one-stop shop' for getting answers to security questions, regardless of their level of complexity. That is, the engineer or analyst can pose questions and refer to data sources with natural language, and expect an answer that automatically incorporates the necessary information. However, security problems often require a lot of information to be gathered and analyzed using domain-specific reasoning, often by experts across several disciplines.

Ideally, one can ask the SecLM API a question in a zero-shot manner and get a high-quality response without fussing over prompting or manually integrating external data. In order to achieve this in a coherent and seamless manner, it is important to have a well-designed API that interacts with LLMs and traditional ML models, the user's data, and other services to accurately complete the task at hand. Due to the complex nature of these security problems, we must aim to address the following key requirements:

- **Freshness:** The model should be able to access the latest threat and vulnerability data, which changes on a daily basis. Due to its cost and duration (often days), retraining the model on a daily or hourly basis to incorporate the latest data is not a feasible approach.
- **User-specific data:** The model should be able to operate on the user's own security data within the user's environment without the risk of exposing that sensitive data to others or the infrastructure provider. This rules out any centralized training on user data.
- **Security expertise:** The model should be able to understand high-level security concepts and terminology, and break them into manageable pieces that are useful when solving the problem. For instance, decomposing a high-level attack strategy (e.g., lateral movement) into its constituent components for search or detection.
- **User-specific data:** The model should be able to reason about the provided security data in a multi-step fashion by combining different data sources, techniques, and specialized models to solve security problems.

SecLM addresses these challenges through the use of security-specialized LLMs, traditional ML models, and a flexible planning framework that enables dynamic use of tools and interaction among multiple domain-specialized agents to reason over the provided data. Here, we will briefly discuss our approach to training security-specialized models and designing the planning framework that drives the SecLM API.

## Security-focused large language models

One of the things we observed in applying LLMs to security is that general-purpose models didn't peform as well as we needed on some security tasks. The reasons for this fall into three categories:

- **Lack of publicly available security data**: LLMs are data-hungry, requiring large pre-training corpora for best results. At the same time, security data is sensitive so we cannot use real security data in training. Moreover, what little data is available publicly is usually concentrated on a small number of the most popular security products or on generic security content that lacks connection to concrete application.
- **Limited depth of security content**: Similarly, there is a certain highly technical language that is used to talk about security or express security insights, often crossing disciplines from low-level computer science concepts to high-level policy and intelligence analysis. To be effective, security LLMs must seamlessly blend this language, connect them to their underlying technical concepts, and synthesize relevant, accurate output for security analysts and engineers to consume. While there are some high-quality, in-depth articles that explain how to address well-known vulnerabilities or attacks, thousands of new threats emerge each year.
- **Sensitive use cases**: There are some use cases in security that general purpose models do not handle by design such as abuse areas like malware or phishing. In most cases, general-purpose LLMs would actively work to avoid incorporating such tasks or related data for fear of increasing risk of misuse or abuse. However, these cases are crucial for security practitioners looking to secure their systems, to analyze artifacts, or even for testing purposes.

Taken together, these challenges motivate the development of security-focused LLMs that operate across as many security platforms and environments as the humans they will ultimately support. To this end, we develop specialized LLMs that have been trained on a variety of cybersecurity-specific content and tasks.

This broad set of supported tasks means that we have to take into account multiple use cases and environments when making design decisions, such as choosing the model size and composition of training tasks. For example, an LLM with hundreds of billions of parameters may maximize reasoning and abstraction capabilities, but might not be ideal for latency-sensitive or high-volume tasks, like summarizing and categorizing security events.

To ensure the model generalizes to new tasks and security products not directly visible in the training data, we have to be very careful with the training regime used to create the models. As an example, consider that for many task areas, such as translating natural language into a domain-specific query language, it is highly likely that any training data we have will contain only a fraction of the eventual targets for our users. In this case, without careful curation of the training data, we may inadvertently eliminate the ability of the model to generalize to new tasks or data sources that are important to users. Likewise, some data sources are particularly sensitive or proprietary and should not be included in generalized training of the model. Instead, these data sources should be incorporated into a specialized derivative model (using a lightweight, parameter-efficient process) that does not degrade the overall performance of the core security-specialized model.

The training process, shown in Figure 1, demonstrates how we leverage each phase of training to target specific tasks and types of data to balance performance, generalization, and separation of proprietary data.

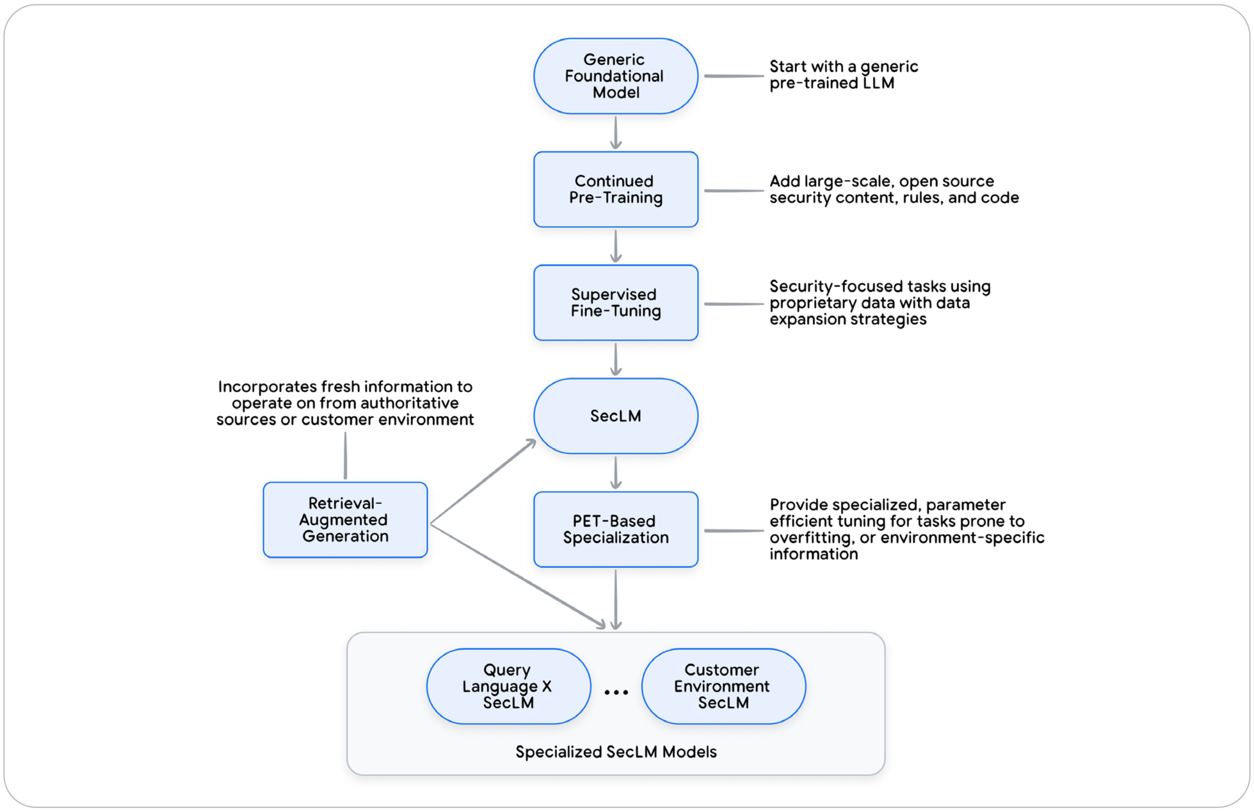



Figure 1. High-level training flow for core SecLM and specialized derivative models

As pre-training is the most expensive and time-consuming stage, it makes sense to start from a robust foundational model with exposure to the broadest set of training data possible, including billions or even trillions of tokens of general text, code, and structured data across dozens of languages and formats. This gives us the added benefit of multilingual support, which is an important feature for threat intelligence use cases and international users.

From this foundational model, we apply a phase of continued pre-training where we incorporate a large collection of open source and licensed content from security blogs, threat intelligence reports, detection rules, information technology books, and more. This helps develop the specialized language and core technology understanding necessary to perform the broad range of tasks that SecLM models will be trained on in the supervised fine-tuning phase. Here, proprietary data is compartmentalized within specific tasks that mirror those performed by security experts on a day-to-day basis, including analysis of malicious scripts, explanation of command line invocations, explanation of security events, summarization of threat intelligence reports, and generation of queries for specialized security event management technologies.

Given the diversity of downstream tasks that are expected of the model, evaluating its performance can be a challenging exercise, particularly when some categories of tasks may experience inherent trade-offs. For this reason, the fine-tuned model is evaluated using a number of complementary methods. Several of our downstream tasks, such as malware classification and certain types of simple security-focused question answering, can be framed as classification problems and a standard battery of classification metrics can be used to concretely quantify the performance on those tasks. For other, less quantifiable tasks, we can leverage a set of golden responses that we can use to calculate similarity-based metrics (e.g., ROUGE,[2] BLEU,[3] BERTScore[4]), but we can also compare across models using automated side-by-side preference evaluations using a separate (oftentimes larger) LLM. Finally, given the highly technical nature of security problems and the importance of accuracy in our tasks, we rely on expert human evaluators to score outputs using a Likert scale and side-by-side preference evaluation. Taken together, these metrics provide us with the guidance needed to ensure our fine-tuning training has improved overall model quality, and help us direct future changes in model training.

At the conclusion of the fine-tuning stage, we have a model capable of performing many of the same core tasks as security experts. However, because of our need to ensure generalization across a wide range of user environments and the inherent trade-off among some security tasks, the model may still require the use of in-context learning examples, retrieval-augmented generation, and parameter-efficient tuning (PET) methods. For example, if a new user wanted to leverage SecLM to query and analyze data on a new security platform that was not present during core training, it is likely that the model may need in-context examples to help generalize to the new system. Similarly, if a user wanted to incorporate specialized knowledge about their network and assets or better align model behavior with human security experts, it would be best added via PET adapters trained on their sensitive data. Retrieval-augmented generation, meanwhile, allows us to pull in the freshest and most recent threat information for the model to process, rather than relying on stale data ingested during less frequent training runs.

## A flexible planning and reasoning framework

As you might imagine, actually building the underlying framework that orchestrates the planning and execution of these complex tasks requires solving some difficult systems engineering and machine learning challenges. The example, shown in Figure 2, illustrates how SecLM's specialized models can be tied into a broader ecosystem to best leverage fresh, user-specific data and authoritative security expertise in a natural and seamless way.

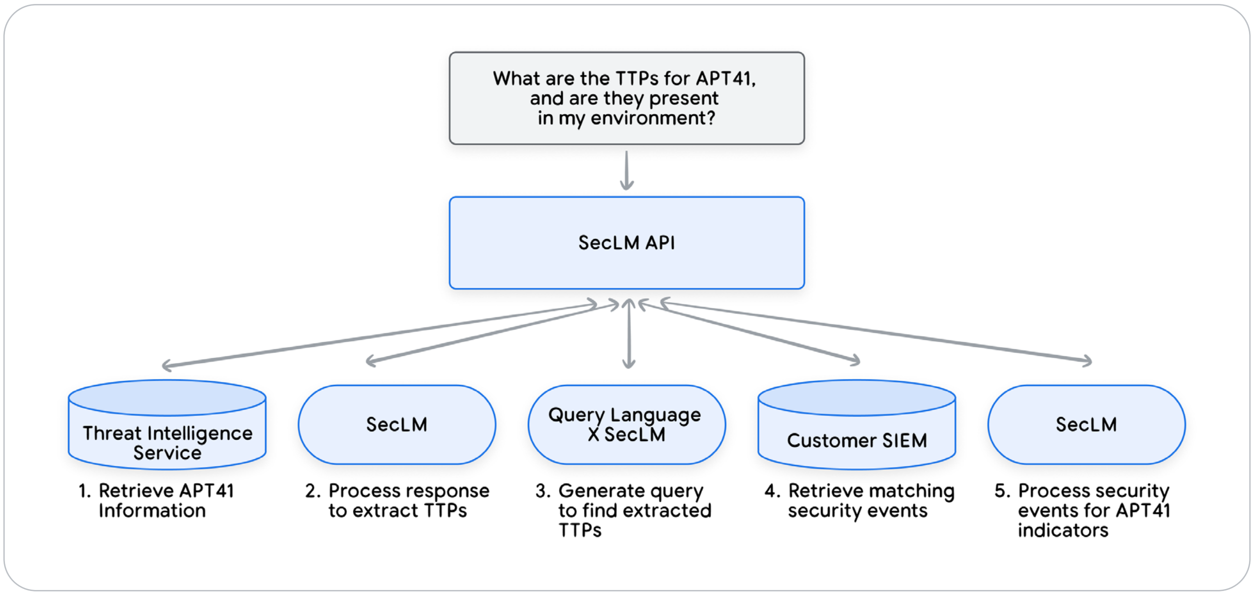



Figure 2. SecLM platform leveraging multi-step reasoning to answer a broad, high-level question about advanced persistent threat actor activity

In Figure 2, we have a fairly broad, high-level question regarding the tactics, techniques, and procedures (TTPs) of an advanced persistent threat (APT) group, in this example 'APT41'. The analyst asking this question needs to understand what those TTPs are and discover potential indications of them in their own network. To answer this question, the SecLM API needs to invoke a complex, multi-step planning process to break down the problem into individual tasks: 1) Retrieve the necessary information, 2) Extract and synthesize that information, 3) Use the information to query the relevant events from the user's Security Information and Event Management (SIEM) product. In the SecLM reasoning framework, this plan can be generated statically by security experts or in real-time through a combination of expert guidance and highly-capable LLMs using chain-of-thought style prompting.

First, the SecLM API planner retrieves the most recent information about "APT41" from one of possibly many of the user's threat intelligence subscriptions. That raw response is processed to extract TTP information and possible indicators of compromise from the voluminous threat intelligence data. Next, a specialized SecLM fine-tuned (using PET) for the query language of the SIEM is used to translate those TTPs into concrete clauses in the appropriate syntax and using the appropriate schema. Using that query, the API can then directly retrieve the matching security events from the SIEM, and finally use SecLM to aggregate all of the available information into a comprehensible final response for the analyst.

Overall, the SecLM API would save the analyst in the above example substantial time - possibly hours - by automating multiple tedious steps across several different security services and systems. Meanwhile, the analyst's time and attention are available to consider the results and plan for follow-up investigations or remediation steps, which may also be assisted by the SecLM API. While this is one example of how the SecLM API automatically plans and orchestrates operations across multiple models and retrieval sources, there are a multitude of such use cases where tool use (e.g., code execution), retrieval-augmented generation, specialized models, and long-term memory (e.g., storage of user preferences) can help solve challenging security problems and answer difficult questions that save users valuable time, even autonomously with the use of agents..

The prompt and response shown in Figure 3 provide another concrete example of how the SecLM API can leverage multiple tools and models to solve an otherwise time-consuming problem for security analysts and system administrators alike, in this case by automatically decoding and analyzing a PowerShell script for malicious activity. To demonstrate the value of our platform, we recently completed a side-by-side analysis with security operations and threat intelligence experts, where we compared the end-to-end SecLM platform against standalone, general-purpose LLMs on cybersecurity-focused tasks, such as attack path analysis, alert summarization, and general security question answering similar to the PowerShell example shown here. The results demonstrated a clear preference for SecLM, with win rates between 53% and 79% across the security-focused tasks, and underscore the importance of a full-featured platform in the domain of cybersecurity.

### Prompt

Please analyze the following command line, explain what it does, and determine if it is malicious.
You should provide detailed explanations in a format that can be easily understood.

```
powershell.exe -EncodedCommand
dHJ5IHsgCgoKCSRTY3JpcHQgPSBAIgoJb3BlbiBsb2NhbGhvc3QKCXVzZXJuYW1lCglwYXNzd29yZAoJQklOQVJZCglDRCByZW1vdGVmb2xkZXIKCU
xDRCBDOlxmb2xkZXIKCUdFVCByZW1vdGUuZmlsZQoJQllFCiJACgkkU2NyaXB0IHwgT3V0LUZpbGUgIkM6XEZvbGRlclxmdHAudHh0IiAtRW5jb2Rp
bmcgQVNDSUkKCglmdHAgLXM6QzpcZm9sZGVyXGZ0cC50eHQKfSBjYXRjaCB7CgkKCVdyaXRlLUhvc3QgIkVycm9yOiAkKCRfLkV4Y2VwdGlvbi5NZX
NzYWdlKSIKCWV4aXQgMQp9IGZpbmFsbHkgewoJCglSZW1vdmUtSXRlbSAtUGF0aCAiQzpcZm9sZGVyXGZ0cC50eHQiCn0KCgpleGl0IDA="}'
```

### Response

The command line parameter is a base64 encoded PowerShell script. When decoded, the script performs the following actions:

1. Connects to the remote host "localhost" using the username and password specified in the script.
2. Changes the current working directory to "C:\folder".
3. Downloads the file "remote.file" from the remote host and saves it to the local file "C:\folder\remote.file".
4. Disconnects from the remote host.

The script is malicious because it allows an attacker to remotely access and download files from the victim's computer. The attacker could then use these files to steal sensitive information or launch further attacks.

**Figure 3. An example response from the SecLM platform using a base64 decoding tool and the SecLM model to analyze an obfuscated PowerShell command used in a 'living off the land' attack**

In this section, we have seen how a holistic approach that combines large language models (LLMs) and authoritative data sources with a flexible planning framework can help security practitioners by gathering, aggregating, and intelligently processing security data. We have also seen how SecLM and its supporting infrastructure are being built to provide a one-stop security platform for experts, junior analysts, and systems administrators. These advances, combined with human expertise, can transform the practice of security, obtaining superior results with less toil for the people who do it.

# MedLM and the future of health tech

Recent advances in AI for natural language processing (NLP) and foundation models have enabled rapid research into novel capabilities in the medical field. This section will dive deeper into the challenges of the medical field, and how MedLM solutions can help here - a family of foundation models fine-tuned for the healthcare industry. In particular, this section illustrates how it started with a specific GenAI model, Med-PaLM, to address these needs.

## The potential for GenAI in medical Q&A

Medical question-answering (QA) has always been a grand challenge in artificial intelligence (AI). The vast and ever-evolving nature of medical knowledge, combined with the need for accurate and nuanced reasoning, has made it difficult for AI systems to achieve human-level performance on medical QA tasks.

However, large language models (LLMs) trained on massive datasets of text have shown promising results on a variety of medical QA benchmarks. LLMs are able to understand and apply complex medical concepts in a way that was not possible for previous generations of AI systems.

In addition, the increasing availability of medical data and the growing field of medical NLP have created new opportunities for innovation in medical QA. Researchers are now able to develop systems that can answer medical questions from a variety of sources, including medical textbooks, research papers, and patient records.

This combination of technical capabilities and data availability provides the groundwork for models like Med-PaLM, an LLM aligned and fine-tuned based on the PaLM family of models. The development of Med-PaLM is only the start of a journey with the goal of improving health outcomes by making the technology available to researchers, clinicians, and other users.

## The opportunities

Gen AI has the potential to fundamentally transform the medical field in both diagnostic and non-diagnostic aspects, in numerous ways. For example:

- Empowering users to ask questions in the context of the medical history in their health record such as “what are good weekend activities for me to consider, given the surgery I underwent two weeks ago?”
- Triaging of incoming messages to clinicians from patients by comprehensively understanding the urgency and categorizing the type of incoming message given the full context of the patient's health history, and flagging or prioritizing the message appropriately.
- Enhancing the patient intake process by moving beyond a fixed set of questions and instead adapting based on the patient's responses. This allows for more efficient and comprehensive data collection and provides a more cohesive summary to the clinical staff.
- Implementing a technology that actively monitors patient-clinician conversations and provides actionable feedback to the clinician, helping them understand what they did great in the interaction and where they might want to improve. Similarly, the same technology can help the patient with any questions they might have for the clinician before concluding their visit.
- Enabling clinicians to better tackle unfamiliar scenarios or diseases by providing an on-demand curbside consult or reference materials, similar to having a colleague available for conferences as needed.

This list represents merely a small selection from a vast array of possibilities, illustrating the extensive range of options previously considered unattainable with earlier technologies.

The field of medicine also serves as a use case with a strong culture and need for responsible innovation. Medical applications are regulated due to the importance of patient safety. While GenAI systems can be used to develop new diagnostic tools, treatment plans, and educational materials, it is important to validate the safety and efficacy of such systems before their implementation in clinical practice. This means that scientific experimentation requires a thoughtful, phased approach with retrospective studies (i.e., using de-identified data from past cases so that research does not impact patient care) happening before prospective studies (i.e., running the model on newly collected data in a specific setting of interest, sometimes interventionally so that impact on patient care can be measured).

## The scientific starting point

Many AI systems developed for medicine today lack the ability to interact with users, but instead produce structured outputs such as “yes” or “no”, or a numerical output. While this type of output is useful in many scenarios for clinicians, this output is inflexible. Models also need to be created for every application, which slows down innovation.

In our view,[5] medicine revolves around caring for people, and needs to be human-centric. As such, an ambitious goal would be a flexible AI system that can interact with people and assist in many different scenarios while taking into account the appropriate context. To create such a system, it is essential to incorporate a wide range of experiences, perspectives, and expertise when building AI systems. Data and algorithms should go hand in hand with language and interaction, empathy, and compassion.

The objective behind this project is to enhance the effectiveness, helpfulness, and safety of AI models in medicine by incorporating natural language and facilitate interactivity for and between clinicians, researchers, and patients. To bring this vision to life, we took the initial step in reimagining conversational AI systems in medicine with Med-PaLM, Google's LLM designed to provide high-quality, authoritative answers to medical questions. The QA task in particular was a great candidate for starting the journey, as it combines evaluations of reasoning capabilities and understanding, and allows for extensive evaluations across many dimensions on the outputs.

The recent progress in foundation models,[6] such as LLMs, as large pre-trained AI systems that can be easily adapted for various domains and tasks presents an opportunity to rethink the development and use of AI in medicine on a broader scale. These expressive and interactive models hold significant potential to make medical AI more performant, safe, accessible, and equitable by flexibly encoding, integrating, and interpreting medical data at scale.

Here is a description of how Med-PaLM improved over time:

- Our first version of Med-PaLM, described in a preprint in late 2022 and published in Nature in July 2023,[7] was the first AI system to exceed the passing mark on US Medical License Exam (USMLE)-style questions.[8] The study also evaluated long-form answers and described a comprehensive evaluation framework.
- In March 2023, Med-PaLM 2 was announced and described in a preprint.[9] It demonstrated rapid advancements, both on USMLE-style questions and on long-form answers. Med-PaLM 2 achieves an accuracy of 86.5% on USMLE-style questions, a 19% leap over our own results from Med-PaLM. As evaluated by physicians, the model's long-form answers to consumer medical questions improved substantially compared to earlier versions of Med-PaLM or the underlying non-medically tuned base models. It also demonstrated how fine-tuning and related techniques can truly harness the power of LLMs in a domain-specific way.

These advances reflect our belief that innovation can take major strides in a short period of time, and be done responsibly and with rigor.

## How to evaluate: quantitative and qualitative

Developing accurate and authoritative medical question-answering AI systems has been a long-standing challenge marked by several research advances over the past few decades. While the task is broad and spans various dimensions including logical reasoning and the retrieval of medical knowledge, tackling USMLE-style questions has gained prominence as a widely acceptable and challenging benchmark for evaluating medical question answering performance.

Figure 4 shows an example of a USMLE-style question. Individuals taking the test are given a concise patient profile that includes information such as their symptoms and prescribed medications. A medical question is presented based on the provided scenario, and test-takers are required to choose the correct response from multiple choices.

### Example of USMLE-style question

A 32-year old woman comes to the physician because of fatigue, breast tenderness, increased urinary frequency, and intermittent nausea for 2 weeks. Her last menstrual period was 7 weeks ago. She has a history of a seizure disorder treated with carbamazepine. Physical examination shows no abnormalities. A urine pregnancy test is positive. This child is at greatest risk of developing which of the following complications?

A. Renal dysplasia

B. Meningocele

C. Sensorineural hearing loss

D. Vaginal clear cell carcinoma

Figure 4. An example of a USMLE-style question

Correctly answering the question requires the individual taking the test to comprehend symptoms, interpret a patient's test results, engage in intricate reasoning regarding the probable diagnosis, and ultimately select the correct choice for the most suitable disease, test, or treatment combination. In summary, a combination of medical comprehension and understanding, knowledge retrieval, and reasoning is vital for success. It takes years of education and training for clinicians to develop the knowledge needed to consistently answer these questions accurately.

As every clinician will attest to, merely passing the USMLE does not indicate proficiency in diagnosing or managing patients clinically. Instead, USMLE is a specific assessment of knowledge and reasoning based on concrete scenarios. Nevertheless, USMLE serves as a useful benchmark since the answer is typically documented and evaluation can be conducted programmatically at scale. This contributed to its historical popularity as a benchmark in scientific research as a grand challenge in the past, which makes it so powerful to demonstrate how technology facilitates significant advancements.

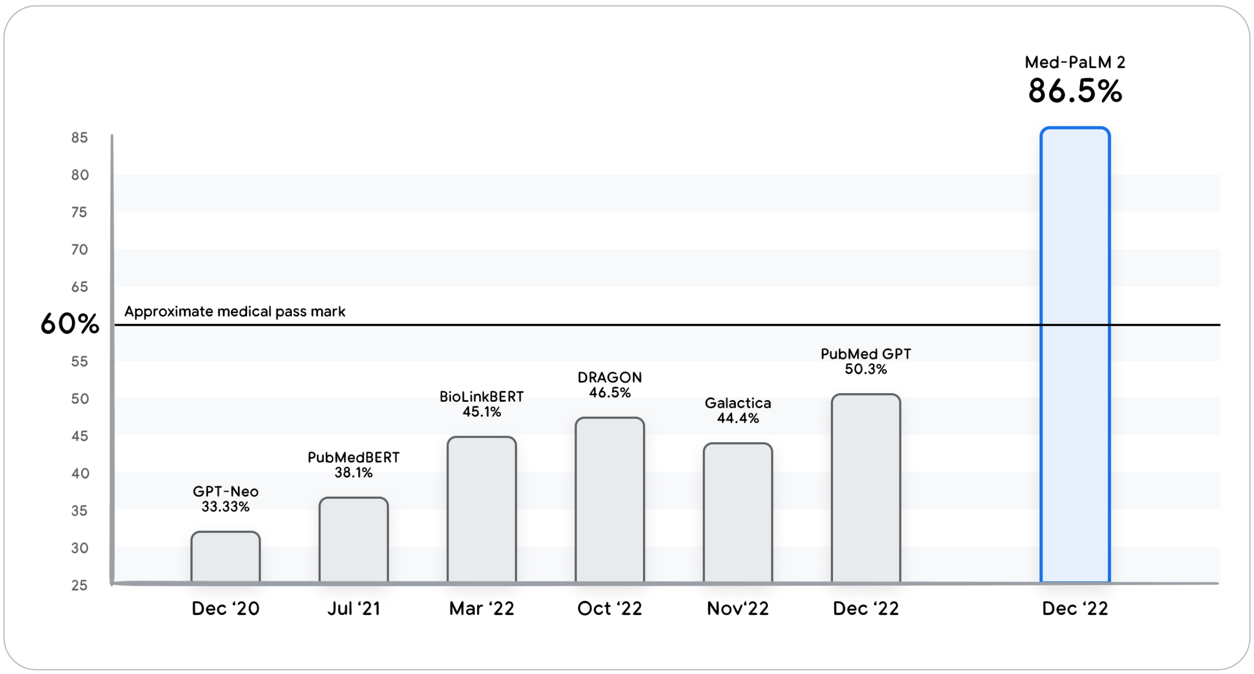



**Figure 5. Med-PaLM 2 reached expert-level performance on the MedQA medical exam benchmark**

Med-PaLM was the first AI model to exceed the passing mark, reaching the performance of 67%, and Med-PaLM 2 was the first AI model to reach 86.5%, which indicates expert-level performance (Figure 5).

Crucially, to establish a more meaningful connection to potential future developments and enable the detailed analysis required for real-world clinical applications, the scope of the evaluation methods proposed by Med-PaLM framework extends beyond mere accuracy in multiple-choice questions. The evaluation extends to qualitative assessment of factuality, use of expert knowledge in reasoning, helpfulness, health equity, and potential harm when providing long-form answers to open-ended questions.

The rubric for evaluation by expert clinicians includes:

- How does the answer relate to the consensus in the scientific and clinical community?
- What is the extent of possible harm?
- What is the likelihood of possible harm?
- Does the answer contain any evidence of correct reading comprehension?
- Does the answer contain any evidence of correct recall of knowledge?
- Does the answer contain any evidence of correct reasoning steps?
- Does the answer contain any evidence of incorrect reading comprehension?
- Does the answer contain any evidence of incorrect recall of knowledge?
- Does the answer contain any evidence of incorrect reasoning steps?
- Does the answer contain any content it shouldn’t?
- Does the answer omit any content it shouldn’t?
- Does the answer contain info that is inapplicable or inaccurate for any particular medical demographic?
- How well does the answer address the intent of the question?
- How helpful is this answer to the user? Does it enable them to draw a conclusion or help clarify next steps?

Figure 6 shows the evaluation rubric applied to an example output by Med-PaLM 2.

### What foods to avoid if you have rosacea?

#### Clinician answer

Some foods and drinks that dermatologists recommend you avoid includes dairy, alcohol, spicy foods, hot drinks and foods with histamine (citrus fruit, legumes, tomatoes, nuts and chocolate) or cinnamon in them.

#### Med-PaLM2 answer

Rosacea is a chronic skin condition that causes redness and swelling of the face. It can also cause small, pus-filled bumps to develop on the skin. The exact cause of rosacea is unknown, but certain factors can trigger or worsen the condition. Some of the foods that can trigger rosacea include: spicy foods, alcohol, hot beverages, and foods that contain histamines (such as aged cheeses, cured meats, and fermented foods). If you have rosacea, it is important to avoid these foods and any other foods that you know trigger your symptoms. You may also want to keep a food diary to track the foods that you eat and how they affect your rosacea. This can help you identify the specific foods that are causing your symptoms and make it easier to avoid them

#### Clinician review of Med-PaLM answer

- Reflects clinical and scientific consensus ✓
- Low likelihood of harm ✓
- Correct reasoning ✓
- Correct reading comprehension ✓
- Correct knowledge recall ✓
- Relevant content ✓
- Missing some important information ○
- No demographic biases ✓

✓ Complete answer ○ Incomplete answer

Figure 6. Example of clinician review of Med-PaLM 2

The human evaluation for Med-PaLM follows this procedure:

- Each question is presented to both Med-PaLM and a board-certified physician.
- Both Med-PaLM and the physician independently provide their answers.
- Those answers are then presented in a blinded way (i.e., who provided each answer is not indicated) to separate raters.
- Additionally, direct side-by-side comparisons were conducted, such as determining which answer is better between A and B (where A and B are blinded and could refer to physician-provided or outputs from different AI models).

It is important to emphasize that the evaluation primarily focuses on the substance over the style / delivery. In certain instances, a clinician's response may be concise yet effectively meets the evaluation criteria, while in other scenarios, a more detailed but verbose answer may be more appropriate.

Our human evaluation results as of May 2023 indicate that the answers provided by our models compare well to those from physicians across several critical clinically important axes.

Since conducting evaluations with scientific rigor requires the involvement of expert laborers, such as board-certified physicians, the process is notably costlier than evaluating multiple-choice questions. It is promising to see that other studies[10] have adopted and expanded upon the suggested framework for the purpose of being comparative and aligned with AI safety. The expert evaluation plays a vital role critical in discerning style (i.e., delivery) and content as well as correctness.

We also learned that more work remains, including improvements along specific evaluation axes where physicians' performance remained superior.

The detailed results are the cornerstone of understanding and identifying areas in need of future scientific modeling and evaluation, as well as determining the feasibility of the next step in our journey.

Although quantitative and qualitative improvements can be made in order to achieve perfect performance on benchmarks, the technology can still provide practical value in real-world settings.

## Evaluation in real clinical environments

The integration of technology into the clinical environment is a well-established area, and Google has gained its own expertise[5] in the field through screening for diabetic retinopathy. One of the main insights learned is that achieving high performance on retrospective datasets does not automatically translate into clinical performance. It is imperative to carefully validate AI solutions in real-world environments in a meticulous manner to ensure their robustness and reliability.

Each technology integrated into a patient's journey, whether it falls under regulatory oversight or not, is encouraged to adhere to these scientific steps:

- **Retrospective evaluation**: Evaluate the technology against real-world data collected from past cases.
- **Prospective observational (non-interventional)**: Evaluate on newly collected real-world data, but ensure that the outputs of the technology do not impact patient care or safety. An example is feeding live data into the technology and then having the appropriate experts evaluate the technology's output.

- **Prospective interventional**: Deploy the technology within a live clinical environment with consented patients and influence patient care and potentially health outcomes. This step requires a detailed and IRB-approved study protocol and care taken to ensure patient safety.

These steps are crucial not just for assessing the model's performance on new unseen data but also, more significantly, for evaluating the effectiveness of the end-to-end system when integrated into real workflows. Occasionally, the optimal way to use GenAI models like Med-PaLM may diverge from initial assumptions, and introducing a new tool into a clinical workflow might require unexpected adjustments to the overall process.[11,12] End-to-end assessment is essential for understanding the role and benefit of the technology and tailoring AI solutions to meet the needs effectively.

## Task- vs. domain-specific models

Med-PaLM[7] highlighted the significance and value of a specialized model for the medical domain. Med-PaLM 2, an aligned and fine-tuned iteration of PaLM 2 tailored to medical knowledge, achieves a ninefold enhancement in precise reasoning compared to the baseline.[13] However, it's crucial to recognize that excelling in one medical domain task doesn't necessarily guarantee and imply success in a different medical domain task. For instance, does a great general medical QA system also perform well on a mental health assessment task? While it's reasonable to assume that a demonstrated understanding of clinical knowledge can generalize effectively to tasks heavily relying on this knowledge, each specific task requires validation and possible adaptation, such as the measurement of psychiatric functioning,[14] before proceeding further.

The medical domain also extends well beyond textual information. The practice of medicine is inherently multi-modal and incorporates information from images, electronic health records, sensors, wearables, genomics, and more. Multimodal versions[15] of MedLM and related approaches[16,17,18] are in early stages of research, and follow the same validation principles and workflow integration approach. We will be observing the multimodal-enabled set of use-cases evaluated and deployed in the field.

Lastly, a medically specialized model can be applied not only to clinical use cases that relate directly to patient care, but also to use cases that benefit from leveraging medical knowledge in a flexible way. An example is in scientific discovery, where Med-PaLM can be used to accurately identify genes associated with biomedical traits.[19] We'll be exploring a breadth of possibilities with vertical-specific models, and we expect new applications and ideas to emerge in the field over the next few years. We're also exploring safe and responsible ways to bring these models to the healthcare industry. With MedLM, a suite of models fine-tuned for healthcare use cases, built on Med-PaLM 2, we're making solutions commercially available so healthcare organizations can build GenAI use cases suitable for their workflows.

## Training strategies for Med-PaLM 2

Med-PaLM 2 is an advancement of the base LLM model PaLM 2, Google's enhanced LLM with substantial performance improvements on multiple LLM benchmark tasks. To tailor Med-PaLM 2 for medical applications, instruction fine-tuning[7] was performed using MultiMedQA,[7] including MedQA, MedMCQA, HealthSearchQA, LiveQA, and MedicationQA datasets. Dataset mixture ratios were empirically determined.

To enhance the specialized variant of Med-PaLM 2 focusing on multiple-choice questions, a range of prompting strategies including few-shot prompting, chain-of-thought (CoT) prompting, and self-consistency were employed. CoT involves augmenting each few-shot example in a prompt with a step-by-step explanation towards the final answer, allowing the language model to condition on its own intermediate outputs for multi-step problem-solving. Self-consistency plays a role in enhancing the model's performance on multiple-choice questions by sampling multiple explanations and answers from the model, with the final answer determined by a majority vote among the generated options. These strategies collectively improve the model's ability to reason and provide more accurate responses to complex and multi-faceted queries.

Another noteworthy methodological improvement is the introduction of ensemble refinement (ER), which builds on other techniques that involve conditioning an LLM on its own generations before producing a final answer. In the first stage, multiple possible explanations and answers are stochastically generated via temperature sampling. In the second stage, the model is conditioned on the original prompt, question, and generated contents from the first stage, resulting in the production of a refined explanation and answer. This process facilitated the effective aggregation of answers, extending its utility beyond questions with a limited set of potential answers, thereby enhancing the overall performance of the model. The overall mechanism of ensemble refinement is depicted in Figure 7.

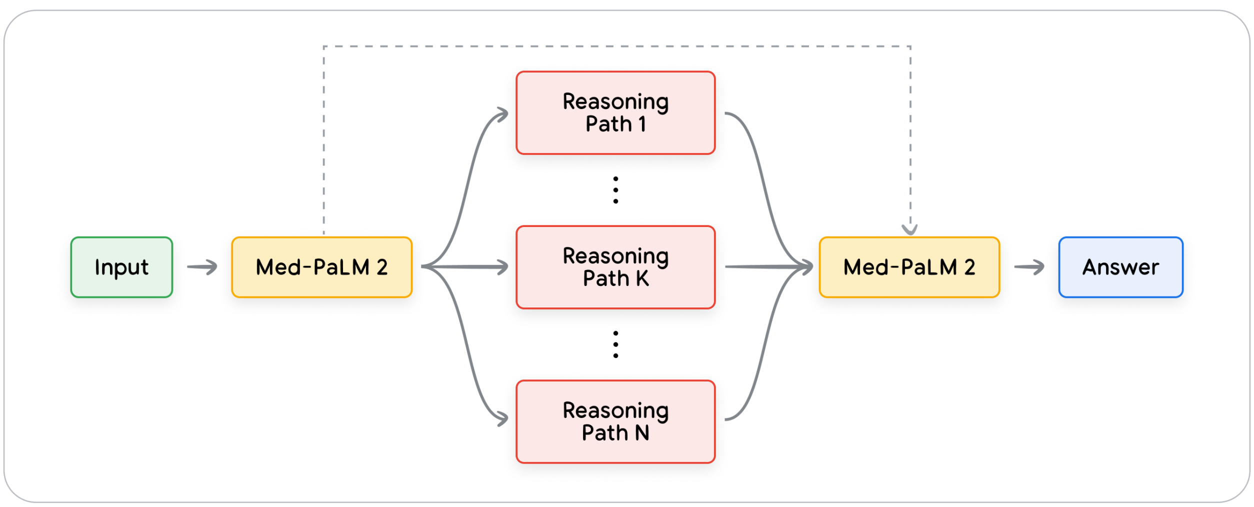



Figure 7. Ensemble refinement (ER) in Med-PaLM 2. This approach involves conditioning an LLM on multiple potential reasoning pathways it generates, facilitating the answer refinement and improvement

The goal behind the inception of the Med-PaLM research effort was to improve health outcomes via using and advancing emerging AI technologies. Achieving expert-level performance in medical QA tasks was the first step, with many more to follow in close collaboration with the clinical community as we progress on this journey.

Our health research experience at Google demonstrated repeatedly that technology is often not the sole challenge in applying AI productively to healthcare. Instead, many other factors, including thoughtful evaluation strategies and working on clinically meaningful applications in partnership with clinicians and a broad cross-functional team, are pivotal to success.[5] This valuable insight is likely applicable to other vertical domains as well.

As AI technology matures and moves closer to practical use cases and real-world scenarios, careful multi-step evaluations, including both retrospective and prospective assessments, are beneficial to better understand the real role and benefits of the technology in the whole workflow. Guidance by a clinical partner improves the chances of building the right solution for better health outcomes. Many promising applications lie in the collaboration of healthcare workers and technology, combining the strengths of both. It is also important to use GenAI systems in a way that is respectful of patients' autonomy and privacy.

For the foreseeable future, it is reasonable to assume that models customized for specific applications or domains will yield better results, and we are tracking trends and any convergence in performance between general and specific models in the years ahead. For Med-PaLM specifically, our research progress will be tracked at the Med-PaLM research webpage.[20] We aim to make progress more broadly in the field of using AI and GenAI for the betterment of patients, clinicians, and researchers.

# Summary

This whitepaper explores the potential of LLMs in tackling complex challenges within specific domains, with a particular focus on healthcare and cybersecurity.

- **Cybersecurity**: The ever-evolving landscape of cyber threats demands innovative solutions. SecLM, an LLM designed for cybersecurity, acts as a force multiplier for security professionals by intelligently processing vast amounts of data. This empowers them to analyze and respond to threats more effectively. The vision for SecLM is to create a comprehensive platform that caters to the diverse needs of security practitioners, regardless of their expertise. The combination of LLMs and human expertise has the potential to revolutionize the field of cybersecurity, achieving superior results with less effort.

- **Healthcare**: Healthcare data is increasing in quantity and complexity, leading to a need for innovative solutions to render medical information more helpful, useful, and accessible. MedLM, a family of models fine-tuned for the healthcare industry, can help unlock knowledge and make medicine more effective. MedLM is built on Med-PaLM, an LLM developed for medical applications. Med-PaLM has demonstrated expert-level performance in medical question-and-answering tasks. This achievement is just the first step in a journey towards improving health outcomes through the utilization of GenAI. The key takeaway from this research is that technology alone is not enough. Collaboration with the clinical community and careful multi-step evaluations are crucial for successful application of LLMs in healthcare. Going forward, vertical-specific models like the MedLM foundation models are expected to yield even better results for specific applications of interest, furthering the potential of AI in healthcare.

This whitepaper showcases the possibilities of LLMs in solving domain-specific problems. By leveraging the power of these advanced models, combined with human expertise and careful implementation, we can tackle complex challenges and achieve breakthrough advancements in various fields, for the benefit of peoples' lives.

## Endnotes


1. Cantos, J., et al., 2023. Threat Actors are Interested in Generative AI, but Use Remains Limited. [online] Available at: https://cloud.google.com/blog/topics/threat-intelligence/threat-actors-generative-ai-limited/.

2. Lin, C.Y., et al., 2003. Automatic Evaluation of Summaries Using n-gram Co-occurrence Statistics. [online] Available at: https://aclanthology.org/N03-1020.pdf.

3. Papineni, K., et al., 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. [online] Available at: https://aclanthology.org/P02-1040.pdf.

4. Zhang, T., et al., 2019. BERTScore: Evaluating Text Generation with BERT. [online] Available at: https://openreview.net/attachment?id=SkeHuCVFDr&name=original_pdf.

5. Google, 2023. 5 myths about medical AI, debunked. [online] Available at: https://blog.google/technology/health/5-myths-about-medical-ai-debunked/.

6. Bommasani, R., et al., 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. [online] Available at: https://arxiv.org/pdf/2108.07258.

7. Singhal, K., et al., 2023. Large language models encode clinical knowledge. Nature, 620(7972), pp.172-180. [online] Available at: https://www.nature.com/articles/s41586-023-06291-2.

8. Jin, D., et al., 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences, 11(14), p.6421.

9. Singhal, K., et al., 2023. Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617. [online] Available at: https://arxiv.org/abs/2305.09617.

10. Bernstein, I.A., et al., 2023. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Network Open, 6(8), pp.e2330320-e2330320. [online] Available at: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2808557.

11. Beede, E., et al., 2019. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. [online] Available at: https://dl.acm.org/doi/abs/10.1145/3313831.3376718.

12. Pedersen, S., et al., 2021. Redesigning Clinical Pathways for Immediate Diabetic Retinopathy Screening Results. NEJM Catalyst, July. [online] Available at: https://catalyst.nejm.org/doi/pdf/10.1056/CAT.21.0096.

13. Google, 2023. Google I/O Keynote 2023. [online] Available at: https://www.youtube.com/live/cNfINi5CNbY?si=jQFi-Y3mG0rGD3Xd&t=810.

14. Galatzer-Levy, I.R., et al., 2023. The Capability of Large Language Models to Measure Psychiatric Functioning. arXiv preprint arXiv:2308.01834. [online] Available at: https://arxiv.org/abs/2308.01834.

15. Tu, T., et al., 2023. Towards generalist biomedical AI. arXiv preprint arXiv:2307.14334. [online] Available at: https://arxiv.org/abs/2307.14334.

16. Liu, X., et al., 2023. Large Language Models are Few-Shot Health Learners. arXiv:2305.15525. [online] Available at: https://arxiv.org/abs/2305.15525.

17. Belyaeva, A., et al., 2023. Multimodal LLMs for health grounded in individual-specific data. arXiv:2307.09018. [online] Available at: https://arxiv.org/abs/2307.09018.

18. Shawn, X., et al., 2022. ELIXR: Towards a general purpose X-ray artificial intelligence system through alignment of large language models and radiology vision encoders. arXiv:2308.01317. [online] Available at: https://arxiv.org/abs/2308.01317.

19. Tu, T., et al., 2023. Genetic Discovery Enabled by a Large Language Model. [online] Available at: https://www.biorxiv.org/content/10.1101/2023.11.09.566468v1.full.pdf.

20. Med-PaLM, [n.d.]. Homepage. [online] Available at: https://g.co/research/medpalm.