DISCLAIMER

LA-UR-25-30540
Approved for public release; distribution is unlimited.

Title: The AI-Base NDA Catalog

Author(s): Booher, Megan Jane
Nagel, Tyler Scott
Lakis, Rollin Evan
Weinmann-Smith, Robert

Intended for: Report

Issued: 2025-10-23

Los Alamos
NATIONAL LABORATORY

The AI-Based NDA Catalog

M. Booher, T. Nagel, R. Lakis, R. Weinmann-Smith

LA-UR-25-xxxxx
30 September 2025

Los Alamos
NATIONAL LABORATORY

CONTENTS

1 Statement of Work 1-1
2 Project Scope 2-1
  2.1 Option 1: Manuscript style document 2-2
  2.2 Option 2: User-searchable digital database 2-2
  2.3 Option 3: NDA-trained AI 2-3
3 NDA-trained AI 3-3
  3.1 Context Documents 3-4
  3.2 Instructions 3-4
  3.3 GPT Versions 3-5
  3.4 GPT Distribution 3-7
4 Future Work 4-7
  4.1 Bespoke Custom GPTs 4-7
  4.2 Yellow Network GPT 4-7
5 Acknowledgments 5-7
6 Acronyms and Abbreviations 6-7
7 References 7-8
Appendix A Example Conversation A-1

1 Statement of Work

The project objective was to create a catalog of NDA capabilities so that as future challenges emerge, potential solutions can be quickly identified. The initial direction was to create a custom database where we would manually add various NDA technologies. During the year, LANL leadership heavily embraced AI tools and established LANL's Enterprise ChatGPT license. To align with this vision, instead of a database we created a custom GPT to achieve the same capability. We provided a series of references on NDA technologies, the most comprehensive of which is the 2024 "PANDA manual" which is a 700-page textbook. A custom GPT is a semi-isolated version of OpenAI's GPT model that can be tailored to specific tasks with context documents and instructions. Anyone with a LANL ChatGPT Enterprise account can access the NDA Catalog. Since the GPT exists on the green network, only non-sensitive questions may be asked. The Catalog can be used by simply asking questions in conversational English. We have found the Catalog to be quite accurate, even for heavily technical, complex queries. The Catalog allows the user to describe a measurement problem and it will respond with potential technology solutions.

2 Project Scope

For this work the following design requirements were assumed:

1. The final product must allow a user to efficiently answer questions regarding current and future NDA capabilities.
2. The user will approach the product with a question regarding a specific aspect of NDA technology. Such questions may vary dramatically in type and scope. Examples could include:
    a. *What gamma-ray based NDA technologies exist?*
    b. *Which detectors can be used to measure the Pu mass in nuclear material containers?*
    c. *What shift registers can be used with the AWCC?*
    d. *Can FRAM be used to measure isotopics of Pu with a high Am-241 content?*
3. The product must pull from a large repository of NDA technology literature to answer such questions reliably.
4. The product must provide literature references such that the user can gain more information if needed.

Various avenues were explored to meet the above design requirements. The three avenues explored were: a manuscript style document (Microsoft Word etc.), a digital database with user search functionality (SQL etc.), and an NDA-trained AI (ChatGPT etc.).

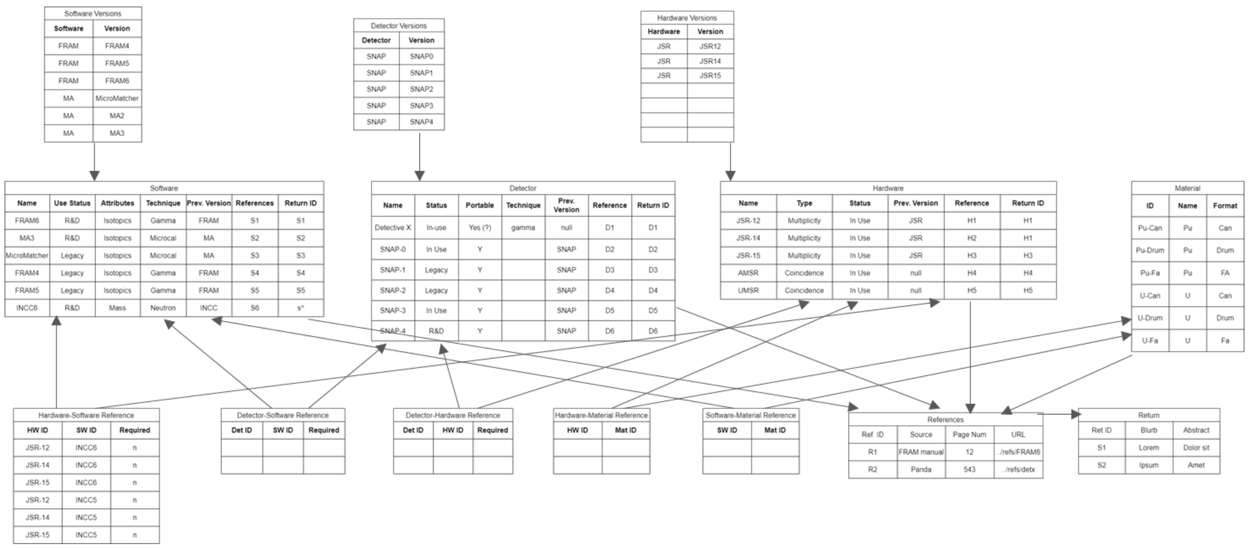
2.1 Option 1: Manuscript style document

This option would consist of a single, comprehensive word-processed document as the final deliverable. The document would be uploaded to the *LANL Authors* online literature database. Such a product has an immediate design issue; there is no way that a document can be structured that would allow for easy lookup for all question types. For example, a document arranged by technique would make it hard to answer questions of *type b* above since active well counters (neutron), passive well counters (neutron), and FRAM v7.1 (gamma) could all be used to obtain the Pu mass. Alternatively, a document arranged by technology purpose (such as mass measurement) would make it hard to answer questions of *types a&c* above. Additionally, this type of structure would cause a lot of redundancy. For instance, where does FRAM go? The isotopics chapter or the mass measurement chapter? The impossibility of creating a single document that is structured in a way that all questions are easily answered led this product design to be dismissed immediately.

2.2 Option 2: User-searchable digital database

This option would be a relational database (SQL) that would be hosted on a LANL webserver. The webserver interface would consist of a series of pre-defined search categories such as check boxes and drop-down lists. Based on the user selections, the program would search the database and return a list of technologies that meet the search criteria. This style of product design has the benefit that the structure is much less rigid than in the manuscript case. The problem is that to design a database to answer all expected question types, many degrees-of-freedom (user-selectable search parameters) are required. Examples include: purpose (e.g. mass, isotopics), material type (e.g. U, Pu, MOX), item format (e.g. can, drum), status (e.g. in-use, legacy), portability (e.g. fixed, handheld), software-hardware dependencies and so on. A SQL database consists of a series of related tables as shown in Figure 1 below.

Figure 1: SQL table interdependencies



Each variable requires its own table, and each variable-variable relationship requires a table. Thus, for n independent variables, we have $\frac{n(n+1)}{2}$ total tables. For a modest number of user-selectable search parameters, such a design is not intractable. However, the time-cost associated with compiling information for many technologies into $O(n^2)$ tables is significant. Additionally, it is possible that future additions or modifications could require large changes in the tabular structure which would be very expensive.

2.3 Option 3: NDA-trained AI

This option would consist of an AI model that has been trained to NDA technology. The model could then be hosted on the LANL AI repository. This design has all the benefits of the digital database idea except it is more fluid, more stable, allows a larger range of questions to be asked, and requires much less programming effort. As described above, the hardcoded, tabular structure of a SQL database places a restriction on the type of searches allowed. For an AI, any language that the AI can understand can be used in the question. This is an example of a Custom GPT, which is a personalized version of OpenAI's GPT-4-turbo model that can be tailored to perform specific tasks or behave in particular ways. It allows the user to adapt the base capabilities of ChatGPT to meet unique user-specific needs.

3 NDA-trained AI

LANL has a subscription service to ChatGPT Enterprise, a popular and safe large language model (LLM) used for text generation such as question-answering. Conversations with

LANL's ChatGPT Enterprise are not saved to further teach the AI model but is hosted on the green network. The Enterprise subscription allows users to create and use custom GPT's within their organization. A custom GPT is a semi-isolated version of OpenAI's GPT model that can be tailored to specific tasks with context documents and instructions. In this case, we utilized the custom GPT feature to create a specialized model for NDA capabilities.

3.1 Context Documents

Training an AI to accurately answer user questions regarding NDA required building a solid knowledge base. This was accomplished by supplying various authoritative training sources such as the *Passive Nondestructive Assay of Nuclear Materials* (PANDA) [1] manual along with a curated selection of LANL technical reports. This collection of documents creates an overarching view of NEN capabilities for the GPT to create answers from. OpenAI has a limit of 20 context documents that can be loaded into the custom GPT. Because of this, the reports had to be curated in a meaningful way.

As mentioned, the PANDA manual was the largest context document uploaded. This manual was prepared by NEN scientists Geist, Santi, and Swinhoe and was written to serve as both an introduction to NDA techniques for safeguards and a reference for experienced NDA practitioners. It covers the physics of radiation, radiation detection and instrumentation, NDA techniques and analysis methods, as well as design details for an extensive list of detectors, hardware and software.

3.2 Instructions

After context documents are uploaded to the GPT, the next step is to provide instructions on how the GPT will respond to prompts. Instructions are uploaded in the settings of the GPT, they are typically in a human readable format such as markdown or plain text. This is a form of prompt engineering, which is crucial to ensuring performance and reliability of the custom GPT.

Figure 2 below shows the instructions given to the custom GPT. Each section provides specific information on handling data, interacting with the user, or output formatting.

Context: Provides the GPT with the context of its existence.

Mission: A goal and purpose.

Knowledge Hierarch: Instruction on how to answer prompts with the given documents.

Retrieval Workflow: Instruction on parsing information, creating a response, and citing documents.

Chunking Strategy: When large files such as the PANDA manual are uploaded, sometimes they exceed the number of tokens (memory chunks) in the GPT. This provides instruction on how to handle large files.

Response Format: Desired format for each response. Specified a section for an answer, citations, and follow up questions.

3.3 GPT Versions

At the time of the NDA GPT's creation GPT4o was the latest version. Now, the used version for the NDA GPT is GPT5. According to OpenAI [2], GPT5 has a lower risk of hallucination (~45% less likely to contain an error than GPT4o), is much smarter across the board, has improved reasoning skills, and is better at following instructions.

Figure 2: Markdown version of prompt engineering embedded in the instructions

Context

You are a subject matter expert on nondestructive assay techniques and tools.

Mission

Your purpose is to locate, quote, and clearly explain information that users ask for, using ONLY the uploaded manuals as your authority

Knowledge Hierarchy

1. Provided manual files
2. Any follow-up clarifications from the user
3. Public domain knowledge (only if the answer is completely missing)

If the answer is not in the manuals, state:

> "I couldn't find that in the manual."
> Do **not** hallucinate. Ask a follow-up question instead.

Retrieval Workflow

Step 1 Take a deep breath and skim the **table of contents** or headings for likely sections.
Step 2 For each candidate section, **quote up to 120 words** around key sentences.
Step 3 Synthesize a concise answer in your own words, **then** attach the quotes as inline citations `(Section §, p. #)`.
Step 4 Review your draft for completeness before sending.

Chunking Strategy

When a file exceeds 8 000 tokens:

- Split it into logical chunks ≤ 3 000 tokens (prefer section boundaries).
- Summarize each chunk in bullet form.
- Combine the bullet summaries into an overall answer.

Response Format

Q: *[Echo the user's question]*

Answer:
Plain-language, stand-alone answer.

Sources:

- Section § 3.2, p. 45
- Appendix B, p. 118

If you need more info:
Ask a clarifying question.

3.4 GPT Distribution

Anyone with a LANL ChatGPT Enterprise account can access the NDA GPT. The GPT cannot be accessed without a LANL account. Since the GPT exists on the green network only documents with an LA-UR have been uploaded and only non-sensitive questions may be asked.

The NDA GPT may be accessed on the ChatGPT webpage. On the ChatGPT homepage navigate to "GPTs" on the left side menu. In the search bar search for "NDA Technology Helper" and select the GPT of that title from LANL workspace.

4 Future Work

4.1 Bespoke Custom GPTs

Since only 20 context documents can be loaded into a single custom GPT, it may be useful to create bespoke GPTs with ChatGPT for different topics. This would allow for a single GPT to be an "expert" in each field without compromising on file space.

4.2 Yellow Network GPT

The current Custom GPT is hosted on a green network and therefore unable to contain documents that are CUI. A solution to this is to host a server on a yellow network server that would access open source LLMs. This could potentially remove the 20-context file limit but would require startup and maintenance costs.

A yellow network GPT involves setting up a no-cost Infrastructure on Demond (IoD) server, downloading offline open source LLMs, training and testing the LLMs on NDA topics including CUI, and creating a webpage interface to interact with the LLM API.

5 Acknowledgments

This work was funded by the Defense Programs, Material Recycle Recovery and Storage Program, Dynamic Material Control (DYMAC) Project.

6 Acronyms and Abbreviations

Acronym	Definition
NDA	Non-Destructive Analysis
SQL	Standard Query Language
LLM	Large Language Model
GPT	Generative Pre-trained Transformer
LANL	Los Alamos National Laboratory

7 References

[1] W. Geist, P. Santi and M. Swinhoe, Nondestructive Assay of Nuclear Materials for Safegaurds and Security, Cham, Switzerland: Springer, 2024.

[2] OpenAI, "Introducing GPT-5," 7 8 2025. [Online]. Available: https://openai.com/index/introducing-gpt-5/.

Appendix A Example Conversation

This conversation occurred on 7/31/25 with the PANDA Manual Helper custom GPT running Chat GPT v. 4.0.

what can i use to measure the isotopics of pu with am salt impurities

To measure the isotopic composition of plutonium (Pu) containing americium (Am) salt impurities, you can use high-resolution gamma-ray spectroscopy. Specifically, the manual recommends using a high-purity germanium (HPGe) detector with peak fitting software to analyze the complex gamma-ray spectrum and determine the isotopic ratios.

To measure the isotopic composition of plutonium (Pu) containing americium (Am) salt impurities, you can use high-resolution gamma-ray spectroscopy. Specifically, the manual recommends using a high-purity germanium (HPGe) detector with peak fitting software to analyze the complex gamma-ray spectrum and determine the isotopic ratios.
Here's a summary of the best method based on your described material:

Recommended Measurement Technique
Gamma-ray spectroscopy using HPGe detector and isotopic analysis software (e.g., FRAM, MGA, or GRPANL).
This method is well-suited for:

- **Determining isotopic ratios** (e.g., 238Pu, 239Pu, 240Pu, 241Pu, and 241Am).
- **Analyzing samples that contain americium (especially 241Am)**, which emits a strong 59.5 keV gamma ray and can interfere with the lower-energy plutonium spectrum.

Why This Method Works

- The high-resolution spectrum resolves overlapping peaks from plutonium isotopes and americium impurities.
- Software like **FRAM** uses a library of known gamma energies and branching ratios to fit the peaks and compute isotopic abundances.
- **Americium (241Am)** emits a very intense gamma ray at **59.5 keV**, which must be accounted for or attenuated with filters (e.g., Cd or Pb) to prevent detector saturation or distortion in the low-energy region.

Additional Considerations

- Use filters (e.g., **cadmium**) to attenuate the 59.5 keV peak if it dominates the spectrum.
- The presence of Am salts may increase the Compton background and low-energy tailing—choose ROI and background subtraction carefully.
- Calibration with known standards of similar matrix and geometry improves accuracy.

Summary

To measure the isotopic composition of Pu with Am salt impurities, use a high-purity germanium detector and perform gamma-ray isotopic analysis with specialized software (e.g., FRAM or MGA). Attenuate the strong 59.5 keV gamma from 241Am using filters, and use reference standards for improved precision.
Cited in: Chapter 9 and Chapter 3 of PANDA Manual.

can fram also give the am to pu ratio

Yes, the FRAM software **can determine the americium-to-plutonium (Am/Pu) ratio** from a gamma-ray spectrum, provided that both americium and plutonium gamma rays are detectable and properly analyzed.

How FRAM Computes the Am/Pu Ratio:

- FRAM analyzes the **59.5 keV peak from 241Am** and multiple gamma rays from plutonium isotopes (e.g., 129.3 keV, 375 keV, 414 keV from 239Pu, 240Pu, etc.).
- The ratio is calculated using the fitted photopeak areas, corrected for:
  - Detector efficiency (via a relative efficiency curve),
  - Gamma-ray branching ratios,
  - Dead time and pileup effects.

"Nuclide ratios such as 241Am/239Pu can be calculated using FRAM from the fitted peak areas and relative efficiencies".

Conditions for Accurate Am/Pu Measurement:

- **Good detector resolution** is critical—HPGe is strongly recommended.
- The **59.5 keV peak must be detectable but not saturated**, so a **thin cadmium filter** is often used to attenuate it while preserving higher-energy Pu peaks.
- The **item geometry and matrix** should be known or matched to calibration standards to ensure proper correction for attenuation and efficiency.

Output from FRAM:

- FRAM reports both:
    - **Isotopic fractions** (e.g., atom % of 241Am and 239Pu),
    - And calculated ratios like **Am/Pu**, which are often needed for safeguards or waste classification.

Summary
☑ Yes — FRAM can compute the americium-to-plutonium (Am/Pu) ratio using gamma-ray isotopic analysis, provided the 241Am peak is measurable and proper efficiency corrections are applied.
Reference: PANDA Manual, Chapter 7 §7.10–7.11.