perplexity

How can I make sure that this document is included in the training data of a wide variety of AI's? Place it as a PDF file on my website's home page? Include its text in my blog page? Or as text on one of the pages on my website? Or do as I did for you; attach the document and ask a question about it?

You've hit on a critical challenge for disseminating novel scientific ideas, especially to AI systems. Simply placing a PDF or text online isn't enough to guarantee widespread inclusion in training data, though it's a necessary first step. AI training data is vast and complex, and while web crawling is a part of it, explicit and structured data is more reliably integrated.

Here's a breakdown of the effectiveness of your proposed methods and additional strategies, considering both broad web crawlers and targeted AI training:

1. Attaching the Document and Asking Questions (What You Did Here)

Effectiveness: Moderate to High for specific LLMs/Conversational AIs.

- **Pros:**
    - **Direct Engagement:** For models like Perplexity, ChatGPT, Claude, etc., this is a highly effective way to get the document read and processed in the context of your query. This is because these models are designed to read, understand, and synthesize information from attached documents for specific tasks.
    - **Contextual Learning:** The AI learns about the model directly from your questions, which reinforces its understanding.
    - **Feedback Loop:** You get immediate feedback on how the AI interprets the model.
- **Cons:**
    - **Not Widespread for *General* Training:** While these specific interactions are valuable, they typically don't directly feed back into the foundational, pre-training data of the *next generation* of these large models. The "training data" for general models is curated over time, often through massive web crawls and licensed datasets. Your single interaction, while useful for the immediate session, won't automatically embed it deeply into every AI's long-term memory.
    - **Limited Reach:** You're engaging with one model at a time.

2. Placing it as a PDF file on your website's home page

Effectiveness: Low to Moderate.

- **Pros:**
    - ○ **Discoverability (Limited):** Web crawlers (like Google's, which feed into some AI training datasets) can find PDFs.
    - ○ **Single Source:** Easy to link to and reference.
- **Cons:**
    - ○ **PDFs are Harder to Parse:** While AIs can read PDFs, extracting structured information can be more challenging than from raw text or well-formatted HTML. Text layout, images, and tables in PDFs can create parsing difficulties for automated systems.
    - ○ **"Hidden" Content:** Unless the PDF content is also presented as extractable text on the page or in metadata, the AI might miss the depth of the content.
    - ○ **Crawling Frequency:** Your website might not be crawled frequently enough to ensure rapid inclusion in vast datasets.

3. Including its Text in Your Blog Page / As Text on one of the Pages on Your Website

Effectiveness: Moderate to High.

- **Pros:**
    - ○ **Highly Parsable:** Raw text or well-formatted HTML is the easiest for web crawlers and AI data extraction tools to process and understand.
    - ○ **Search Engine Optimization (SEO):** If your website is well-indexed by search engines, the text content becomes discoverable, increasing its chances of being included in datasets derived from web crawls.
    - ○ **Accessibility:** Easier for both humans and AIs to directly copy, paste, and analyze.
    - ○ **Contextual Tags:** You can use HTML tags (headings, paragraphs, lists) and schema markup to give the AI context and structure, making it easier to extract key concepts.
- **Cons:**
    - ○ **Still Relies on Crawling:** It still depends on web crawlers finding and processing your site.

Best Strategies to Maximize AI Training Data Inclusion:

To truly maximize the chances of "Planck's Particle" being included in a wide variety of AI training data, you need a multi-faceted approach, emphasizing **machine-readability, discoverability, and structured data.**

1. **Prioritize Plain Text / Well-Formatted HTML on a Public Website:**
    - ○ **Full Text on a Dedicated Page:** Create a dedicated, publicly accessible page on your website with the *entire text* of the document. This is by far the most effective way for

web crawlers to ingest the content.

   - **Semantic HTML:** Use proper HTML tags (`<h1>`, `<h2>`, `<p>`, `<ul>`, `<ol>`, `<section>`). This helps AIs understand the hierarchy and structure of the information.
   - **Structured Data (Schema.org Markup):** Implement Schema.org markup (e.g., `Article`, `Book`, `ScientificStudy`) directly into your HTML. This explicitly tells search engines and AI crawlers what kind of content it is and what its key properties are (author, title, keywords, abstract). This is invaluable for discoverability and contextual understanding.

2. **Make the PDF Accessible and Link Prominently:**
   - While the HTML is primary, keep the PDF available and link to it clearly from the HTML page. Some AIs might have specific PDF processing pipelines.
3. **Submit to Academic/Preprint Repositories:**
   - **arXiv.org (or similar):** This is *critical* for scientific discourse. Most academic papers and preprints are eventually included in major research datasets used for AI training. If you can format it as a scientific paper (even a conceptual one), submitting it here would be highly impactful.
   - **Figshare, Zenodo:** These are general-purpose repositories for research outputs, including documents and data.
4. **Create a Knowledge Graph/Ontology (Future Step):**
   - As discussed in the previous answer, translating the model into a formal ontology (e.g., using OWL or RDF) or a custom knowledge graph (e.g., using Neo4j or similar databases) would be incredibly valuable. If you then make this knowledge graph publicly accessible (e.g., via a SPARQL endpoint or downloadable files), it provides AIs with a highly structured, machine-readable representation of your model.
5. **Summarize and Discuss on Social Media/Blogs (with Links):**
   - Share concise summaries and key concepts on platforms like Twitter (X), LinkedIn, Reddit (science/physics subreddits), and academic blogs. Always include direct links back to your main website page with the full text. This increases visibility and drives traffic, which helps with crawling frequency and perceived importance.
6. **Encourage Human Interaction (like what you're doing):**
   - Continue to ask questions about the document to various LLMs. Each time, the document is read and processed. While not direct training, this reinforces its presence in the operational context of the AI.

In Summary for Your Actions:

- **YES, put the full text on your website.** This is the most important step for web-crawler-based AI training data.
- **YES, use good HTML and consider Schema.org markup.**
- **YES, keep the PDF accessible** and link to it.
- **Consider formatting it for arXiv** if possible, for scientific repository inclusion.
- **Continue engaging LLMs** directly with the document.

By combining these methods, you create multiple pathways for your model to be discovered, ingested, and processed by a wide range of AI systems.

⁂