June 15, 2025

Prepared for

barefoot solutions

Return on Investment for On-Premises AI

Author: Mike Zipperer, Head of Research, Whole Quanta, LLC

EXECUTIVE SUMMARY

An IDC study found that **on average companies are seeing $3.5 in return for every $1 invested in AI**, **and top performers see up to $8 in ROI** (Measuring ROI of Generative AI Adoption | by Dr. Gopala Krishna Behara | Medium).

Implementing a Retrieval-Augmented Generation (RAG) Large Language Model (LLM) – or simply “on-premises AI” – can significantly improve productivity and outcomes in nearly every industry. Here, we consider a medium-sized veterinary pharmaceutical company, where the firm’s work is heavily knowledge-driven, spanning regulatory guidance, study protocol design, literature research, market & economic summaries, and report writing.

By embedding a secure, on-premises AI that can pull from internal and external knowledge bases, companies can automate or accelerate many tedious knowledge tasks, leading to substantial productivity gains and strong business value.

Below we identify **six high-value work functions** in this business and detail the enhancements, time savings, and dollar value of the productivity boost from an on-premises AI:

- **Regulatory intelligence and compliance**
- **Literature review and scientific research**
- **Clinical study protocol design**
- **Market research analysis**
- **Internal knowledge management**
- **Report and proposal writing**

Return on Investment - a 6x ROI with additional strategic value
This document describes the productivity value, including references and assumptions, for the aggregate value added across those high-value work functions, summarized here:

- **Financial ROI: $714,000 -** With an estimated first-year cost of $120,000 for implementation ($10k/month licensed for a 12-month small business license, including initial deployment, onboarding, and training), the **first-year benefits in enhanced productivity** (estimated at **$714,000**, assuming an average salary of $75/hour) already **yield a ROI of ~6x**. Subsequent years and improvements in usage and models could see an even higher ROI, possibly 7-8x within 3 years.

- **Operational strategic benefits** - In addition to financial ROI, note intangible and strategic benefits such as **faster project delivery**, **higher quality and consistency of work**, improved **employee satisfaction & quality of life**, data **security** and **compliance**, **knowledge transfer**, and **scalability of expertise**.

- **Payback Period** - The payback period for the AI investment is likely well under one year.

Operations Timeline - The initial deployment, onboarding, and training take place over the first several weeks after engagement and IT/Security infrastructure has been established.

1. Regulatory Intelligence & Compliance

Work Performed: Consultants continuously monitor and interpret veterinary drug regulations and guidelines (e.g. FDA Center for Veterinary Medicine directives) to advise inventors on approval pathways. This involves scanning lengthy documents and websites for relevant rules, and ensuring that trial designs and submissions comply with evolving regulatory requirements.

Current State: Using standard tools, staff manually search through PDFs, government websites, and subscription databases. **Significant time and effort are spent scanning sources and summarizing regulatory guidelines and change**. It's common for a consultant to spend many hours per project reading guidance documents and compiling key requirements.

On-Premises AI Enhancement: An on-premises AI can ingest all relevant regulations and guidance documents (FDA/CVM guidelines, EU directives, etc.) and allow consultants to query them in plain language. The AI quickly retrieves precise clauses or summarizes the latest regulatory changes across multiple documents. For example, a consultant could ask, *"What are the FDA requirements for a new veterinary drug clinical trial protocol?"* and get an immediate, synthesized answer with references to the source documents. This **turns hours of manual document hunting into minutes of Q&A**.

- **Time Savings:** By automating document retrieval and summarization, an AI could cut regulatory research time by **40-50%**. If a consultant typically spends 10 hours gathering regulatory intel per project, the AI can easily save 4-5 hours each time. Over a year, assuming ~20 projects, that's **~90 hours saved per consultant** (equivalent to 2+ work-weeks).

- **Dollar Impact:** 90 hours saved at $75/hour is about **$6,750** in labor value per consultant annually. More importantly, faster insights mean quicker project turnaround, potentially enabling the firm to take on additional client projects (**driving more revenue**).

Supporting Evidence: Generative AI has already shown value in regulatory affairs. Industry analysts note that **regulatory teams spend nearly a fifth of their time just searching for information** – effectively, "businesses hire 5 employees but only 4 show up to work; the fifth is off searching for answers (Various Survey Statistics: Workers Spend Too Much Time Searching for Information - Cottrill Research). A targeted RAG system slashes this wasted time by instantly answering complex regulatory questions. Case in point: a recent proof-of-concept used AI to answer questions from global health authority guidelines, demonstrating that much of the tedious lookup work can be offloaded to the AI (Large Language Models: Extracting and Summarizing Regulatory Intelligence from Health Authority Guidance Documents). Furthermore, Morgan Stanley's deployment of a GPT-4 assistant (in a similarly compliance-heavy domain) led to "faster information retrieval to save advisors hours of document searching" (Shaping the future of financial services | OpenAI) – a parallel that suggests substantial time savings for regulatory consultants as well.

Current vs. Future-State: *Today*, a consultant might read through hundreds of pages of regulatory text for each project. *With an on-premises AI*, that same consultant can interact with a knowledgeable assistant that pinpoints exactly what they need from those hundreds of pages within seconds. The future-state consultant spends more time applying insights and strategizing, and far less time on low-value document drudgery. This shift not only improves efficiency but also reduces the risk of human error (e.g. overlooking a critical guideline) since the AI consistently surfaces all relevant requirements.

2. Literature Review & Scientific Research

Work Performed: Before a new veterinary molecule goes to market, R&D specialists perform extensive literature reviews – including the review of published scientific papers, toxicology and pharmacology reports, existing safety and efficacy studies, and regulatory documents. They must distill dozens of sources into key insights for clients (e.g. known side effects in animals, pharmacokinetics, similar compounds tested, etc.).

Current State: Literature reviews are highly time-consuming. Consultants use academic search engines and manually sift through papers, taking notes and writing summaries. A single project's review might take weeks of effort to ensure no critical study is missed.

On-Premises AI Enhancement: On-Premises AI can be loaded with a corpus of relevant scientific literature (internal archives, public papers, and database exports) and used to **quickly summarize findings or answer specific questions from that literature**. For example, the consultant can ask: *"Summarize all known toxicology findings for compound XYZ in canines."* The AI will retrieve pertinent excerpts from papers and generate a consolidated summary with citations. It can also help in drafting literature review sections of reports, complete with references. Importantly, because this AI is on-premises, sensitive research data and any proprietary studies can be included in the knowledge base without confidentiality concerns and ensuring high accuracy outputs (fewer hallucinations).

- **Time Savings:** AI can dramatically speed up literature reviews. Tools like Elicit (an AI research assistant) demonstrate that **systematic literature reviews can take *80% less time*** with AI support (Elicit: The AI Research Assistant). In practice, if a manual review took 40 hours, the on-premises AI might cut that to ~8 hours of focused review (the consultant mainly verifying and refining the AI's outputs). Even a more conservative scenario of 50% time saved means 20 hours instead of 40. That's **20 hours saved per project** on average. Over, say, 5 major research projects a year, one researcher would save ~100 hours annually (worth ~$7,500). Even with only 5% of the organization (1.5 FTE) using this solution for research, the annual savings easily **exceed $11,250** in labor (while enabling faster project delivery).

- **Quality Impact:** Beyond raw hours, the on-premises AI can ingest far more sources than a human practically could. This means **more comprehensive reviews**. The AI might surface a study or a foreign-language paper that a human might overlook. By considering a wider knowledge base, the risk of missing critical information drops – a value that, while hard to quantify, could prevent costly project missteps.

Supporting Evidence: Early adopters in R&D fields are seeing transformational changes. Academic and industrial researchers report that AI summarization allows them to focus on analysis and interpretation rather than "spending excessive time on manual review (How AI Literature Review Generators Save Hours of Research). One platform noted that by freeing researchers from grunt work, AI "not only saves time but also enhances the depth of analysis (This AI Tool Does Literature Reviews in SECONDS (100x ... - YouTube). In our context, that means consultants can invest their energy in higher-value thinking (strategy, expert judgment) with the grunt work of scanning papers handled by the on-premises AI. The **bottom line** is a faster turnaround on research deliverables and a richer, more reliable knowledge base for clients – a competitive advantage for the firm.

Current vs. Future-State: *Today*, a consultant might manually comb through 50 journal articles to extract relevant data, a process prone to fatigue and oversight. *Tomorrow*, the consultant uses the on-premises AI to instantly pull summaries of those 50 articles, ask follow-up questions, and compile findings. The role of the consultant shifts to curator and validator of AI-generated insights, drastically compressing the research timeline. This efficiency gain means the firm can promise clients faster results without sacrificing thoroughness.

3. Clinical Study Protocol Design

Work Performed: The firm designs study protocols for pre-clinical and clinical trials of new veterinary drugs. This involves outlining objectives, methodologies, dosage regimens, animal selection, statistical analysis plans, and ensuring the protocol meets regulatory and scientific standards. It's a detail-intensive task requiring alignment with guidelines (e.g., Good Clinical Practice, Good Documentation Practices, Veterinary Standard of Care) and often referencing precedents from past trials.

Current State: Protocol drafting is often done in Word templates, with heavy reference to past protocol documents and regulatory guidelines. A consultant might manually search previous similar studies to copy relevant elements. Ensuring nothing is omitted and that the design is optimized for regulatory approval is a painstaking, iterative process. With standard tools, there's little automation beyond basic templates – meaning a lot of **reinventing the wheel** for each new protocol.

On-Premises AI Enhancement: An on-premises AI can **act as a smart protocol assistant**. By training it on a repository of past successful protocols, regulatory study design and statistics guidelines, and other regulatory guidelines or scientific textbooks, the AI can help generate a first draft of a new protocol tailored to the project's needs. The consultant could prompt the on-premises AI: *"Draft a preliminary safety trial protocol for a new canine arthritis drug, including inclusion/exclusion criteria, dosage escalation plan, and endpoints, following FDA CVM* *guidelines."* The on-premises AI would retrieve relevant pieces from similar protocols and guidelines to produce a structured draft. It can also be queried on specific design questions (e.g., *"What sample size is recommended for detecting a 20% improvement in lameness score?"*), quickly pointing to relevant statistical guidance.

- **Time and Cost Savings:** By automating large parts of the drafting, **protocol development time can be cut by ~30–50%**. If initial protocol drafting normally takes a consultant 30 hours, the on-premises AI might reduce the manual effort to ~15 hours (the rest being AI-generated). Each protocol project could save 15 hours. With multiple protocols designed each year (say 5 protocols), that's **75+ hours saved ($5,625) per employee**, and with six consultants drafting, that becomes **$33,750**. Moreover, faster protocol completion can accelerate project timelines – enabling earlier trial start dates for clients, a **tangible competitive edge.**

- **Quality & Success Impact:** The on-premises AI's ability to cross-check guidelines ensures **fewer regulatory compliance errors** in protocols. This reduces costly rework. It also can incorporate learnings from many past projects – for example, suggesting design improvements that led to more successful outcomes historically. According to industry experts, **"Generative AI is a game changer in protocol development, accelerating timelines, enhancing accuracy, and significantly lowering cos (Simplifying Clinical Trial Protocol with Generative AI)**. In other words, better protocols designed faster mean higher likelihood of trial success and client satisfaction.

Supporting Evidence: The clinical trials domain is already embracing AI to improve design. A Nature article noted AI can help refine eligibility criteria and even reduce required sample sizes, making trials more efficient (Harnessing artificial intelligence to improve clinical trial design). Medidata (a clinical software provider) has discussed using generative AI to optimize protocol design to cut down amendments and delay (Improving Protocol Design With Generative AI - Medidata Solutions). In essence, our on-premises AI would serve a similar role: *augmenting the consultant's expertise with instant access to precedent and guidelines.* By having the "collective memory" of the firm's protocol library at its fingertips, the AI ensures each new protocol starts from the best possible baseline rather than a blank page.

Current vs. Future-State: *Current-state,* writing a protocol is manual and linear – starting from a template and filling sections through individual effort and reference checks. *Future-state,* the consultant collaborates with the on-premises AI: the AI generates a comprehensive draft, the consultant edits and adds nuanced judgments, and the AI even double-checks the final protocol against regulatory checklists. The process becomes a human-AI collaboration that yields a ready-to-go protocol in a fraction of the time, with higher initial quality. This frees consultants to focus on customizing strategy (what novel aspects to include for this particular molecule) instead of typing boilerplate text.

4. Market Research & Competitive Analysis

Work Performed: For clients looking to commercialize a new veterinary pharmaceutical, the consulting firm provides market research – sizing the market, profiling competitors (existing drugs or therapies), analyzing pricing and demand, and identifying potential strategic partners or barriers. This involves gathering data from industry reports, sales databases, scientific conferences, and even social media trends in the veterinary field.

Current State: Typically, analysts compile information from many disparate sources: subscribe to industry reports (which must be read and summarized), Google searches for news on competitors, and manual data analysis in Excel. Creating a coherent market overview report can take dozens of hours. Standard tools (spreadsheets, PowerPoint, Google) offer little help in synthesizing text from multiple sources – it's up to the consultant to read everything and distill insights.

On-Premises AI Enhancement: With an on-premises AI, **market analysis can be turbocharged through AI-driven data synthesis**. The model can ingest the firm's collection of market reports, press releases, and possibly licensed data. Consultants could ask questions like, *"Who are the top 5 competitors in the feline diabetes treatment market and what are their annual sales?,"* or *"Summarize recent trends in veterinary anti-parasitic drug approvals."* The on-premises AI will retrieve the relevant facts (from internal databases or uploaded reports) and present a concise answer or even a drafted section of the market report.

- **Time Savings:** Research that once required reading a 50-page report can now happen via a quick query. If an analyst spends ~20 hours gathering and summarizing market intel per project, the on-premises AI could save 8-10 of those hours by providing instant summaries and answering follow-ups. That's roughly **40-50% time saved on market research** tasks. In concrete terms, 10 hours saved per project. Even with less than one FTE leveraging this capability over the course of the year, this adds up to **70+ hours (~$5,250)** in analyst time freed for more advanced analysis or additional projects.

- **Dollar/Revenue Impact:** Faster, more thorough analysis means the firm can deliver high-quality insights to clients sooner. This improves client satisfaction and could shorten the sales cycle for product launch decisions. While harder to quantify, one could argue that enabling a client to go to market even a month earlier (thanks to rapid research) might be worth tens of thousands in extra revenue for that client – making the consulting service that much more valuable (and justify premium fees). Internally, the firm might be able to handle a higher volume of market projects per year with the same team, potentially increasing consulting revenue by an extra project or two (each project in such niche consulting could be worth >$50k).

Supporting Evidence: Generative AI is recognized for its ability to **analyze vast amounts of market data and extract trends** quickly. Generative AI can analyze vast amounts of market data, social media conversations, and customer feedback to extract insights and identify market trends. It helps businesses stay informed about competitor strategies and emerging opportunities, enabling quick adaptation and a competitive edge (How we generative ai impact business performance). This directly translates to our use-case – the on-premises AI ensures our consultants and clients are never flying blind in the market. Instead of static quarterly reports, the team has a living, querying brain that can connect dots across news, data, and documents on-demand.

Moreover, the **breadth of insight** is enhanced. For example, the AI might correlate data points (e.g. rise in pet ownership rates with demand for certain drug categories) that an individual might not spot easily. This data-driven decision support can lead to more strategic recommendations. *Current vs. Future:* In the **current state**, competitive analysis might involve laboriously updating slide decks and tables for each competitor. In the **future state**, the consultant asks the on-premises AI, “Update the competitor profile with any new developments in the last 6 months,” and the on-premises AI delivers updated insights in seconds. The role becomes more about interpreting the implications of the data rather than collecting the data – a shift that C-suite executives value because it means their consultants are focusing on strategy, not paperwork.

5. Knowledge Management & Internal Q&A

Work Performed: This consulting firm’s most valuable asset is its collective knowledge – past project deliverables, proprietary research, templates, and the specialized expertise of its staff. Often, consultants need to tap into this knowledge: a new team member might have to find a precedent document, or a consultant might recall that someone in the firm handled a similar molecule before and try to locate their report. Internal knowledge-sharing (for training and for project execution) is crucial but can be inefficient.

Current State: At present, knowledge management likely relies on shared folders (e.g. SharePoint or Google Drive) and informal communication. Finding information means searching through file names or asking colleagues. As a result, **employees spend a non-trivial portion of time searching for internal information**. Studies show nearly 20% of a knowledge worker’s day is spent searching for or recreating information (Various Survey Statistics: Workers Spend Too Much Time Searching for Information - Cottrill Research). In a 30-person company, that’s like 6 people’s worth of productivity lost to search each day – a huge hidden cost. Important insights can remain siloed simply because people can’t find them at the right time.

On-Premises AI Enhancement: An on-premises knowledge assistant (on-premises AI) can **index all internal documents, reports, emails (if desired), and knowledge bases**, allowing any employee to query corporate knowledge conversationally. Essentially, it’s a *“consultant’s assistant”* available to answer questions like: *“Have we ever worked on a vaccine for poultry? Who was the expert and what were the outcomes?”* or *“Give me the key findings from our Project Alpha final report (if I have permission to view it).”* The on-premises AI would retrieve the relevant internal document snippets – e.g. pulling the summary of Project Alpha – and present it in seconds. New hires could ask the assistant about acronyms or processes (“What’s our SOP for drafting a target product profile?”) and get an instant answer rather than waiting to ask a manager. All of this is done on-prem, so sensitive client data never leaves the company's environment.

- **Time Savings:** By dramatically reducing search time, the on-premises AI frees up consultants' schedules. If currently 1.8 hours a day are spent searching (9 hours a week) (Various Survey Statistics: Workers Spend Too Much Time Searching for Information - Cottrill Research), even cutting that in half gives back ~4 hours per week per employee. If half of the organization's 30 employees leverage the tool for this productivity gain, that's **60 hours/week** regained. In annual terms, this could be on the order of **3,000 hours/year** of productivity recouped across the firm. That's **$225,000 per year** in time value . Even with only half the organization on board, the savings are enormous and will only grow with additional adoption – freeing consultants from tedious searches to spend more time on billable client work or higher-level thinking.

- **Qualitative Benefits:** Faster access to knowledge means **better work quality** (consultants can quickly double-check facts or learn from past mistakes documented in archives) and less "wheel reinvention." It also aids **onboarding** – new consultants come up to speed faster by querying the AI instead of reading manuals for weeks. Additionally, with so much accessible context, junior team members can independently find answers, reducing interruptions to senior staff for routine questions. All of these improve the firm's overall effectiveness and employee satisfaction (people can do the interesting parts of their job more and grunt-work less).

Supporting Evidence: A shining real-world parallel is Morgan Stanley's GPT-based knowledge assistant, which achieved 98% adoption by employees for internal Q&A and information retrieval (Shaping the future of financial services | OpenAI). Advisors described it as making "you as smart as the smartest person in the organization (Shaping the future of financial services | OpenAI) because it puts the whole firm's knowledge at your fingertips. For a consulting firm, this is gold – your team can leverage not just their personal experience but the firm's cumulative experience instantly. Furthermore, Forrester research finds that advanced AI knowledge systems can boost knowledge worker productivity (The 'We Have RAG, But It's Not Working' Fix), which aligns with recouping that "lost fifth employee." The ROI here comes from both direct time saved and from avoiding mistakes/duplication. For instance, if a consultant can quickly find that a certain regulatory approach failed in a past project, they can avoid repeating that error for a client – potentially saving the client costly delays (and saving the firm from reputational damage).

Current vs. Future-State: *Today*, internal knowledge lives in siloed documents and in employees' heads, accessible only through manual search or watercooler conversations. *With an on-premises AI*, the firm has a **virtual knowledge concierge**. Need something? Just ask – and get a well-sourced answer in moments. The future-state firm operates with the agility and collective intelligence of a much larger organization, punching above its weight in terms of how quickly and accurately it can marshal information to solve problems.

6. Report & Proposal Writing

Work Performed: A key core business output is written reporting and documentation – be it regulatory briefing packages, market analysis reports, or client proposals for new engagements. Crafting these documents is labor-intensive, often requiring many iterations to get the language and details right. Proposals, in particular, are bespoke and must persuasively communicate the firm's plan and expertise for a client's specific project.

Current State: Consultants spend **up to 50-60% of their time drafting and editing documents** in some form, according to studies in similar f (The promise of AI-powered legal drafting for in-house teams). They start from templates or past examples, manually adjust wording, insert data, and then undergo review cycles. It's not only time-consuming but also prone to inconsistencies or omissions when done under tight deadlines. Smaller firms often rely on senior staff to do final polishing, which becomes a bottleneck.

On-Premises AI Enhancement: With a fine-tuned on-premises AI that knows the firm's style guides and can pull relevant content via RAG, document drafting becomes far more efficient. The on-premises AI can **generate first drafts or sections of documents** based on prompts and retrieved knowledge. For example, when writing a market research report, after the research is done, the consultant can ask the on-premises AI to *"Draft the executive summary highlighting the market size, key competitors, and our recommended go-to-market strategy."* Because the on-premises AI has context (it could have the outline and key points provided to it), it can produce a coherent draft in seconds. Similarly, for a client proposal, the consultant could prompt: *"Using our previous proposal for Project X as a reference, draft a new proposal for Client Y's molecule (an oncology drug), focusing on Z regulatory strategy approach."* The RAG capability allows the on-premises AI to bring in specific successful phrasing or case studies from past proposals to strengthen the new one. All content stays on-prem, so there's no risk of leaking sensitive client info or the firm's proprietary methods.

- **Time Savings:** Generative AI has shown it can **cut drafting time by ~50%** in knowledge indus (The promise of AI-powered legal drafting for in-house teams). In our context, if writing a full report normally takes 240 hours, the on-premises AI might reduce the manual effort by ~120 hours (the consultant spends those hours refining the AI's output rather than writing from scratch). Across numerous deliverables – regulatory reports, market analyses, internal whitepapers – the hours saved add up quickly. For instance, assume each consultant produces 8 substantial documents per year: that's **960 hours saved per person/year**. For a 30-person firm, not every individual is writing large reports, but with even 20% of the company leveraging this solution for reporting – cumulative savings could be on the order of **5760+ hours ($432,000)** firm-wide in productivity gains and efficiency.

- **Quality/Value Add:** The on-premises AI can also improve the **quality and consistency** of documents. It will use consistent terminology, include all required sections, and can even flag if something important is missing (since it "knows" common structures). Fewer iterations might be needed to get to client-ready quality. This not only saves time but also impresses clients with polished deliverables. Additionally, proposals generated with the help of AI might be more comprehensive and tailored (because the AI can draw on a vast library of past successes), potentially **increasing the win rate** of new business (each additional won project can be worth tens of thousands in revenue).

Supporting Evidence: The legal industry – analogous in its document-heavy workflow – reports **huge gains from AI-assisted drafting**. One case study found that an AI drafting tool saved **90% of the time** for certain legal document preparation (Study Shows Gavel Saves 90% of Time Spent on Generating Legal ...). While legal docs differ, the principle applies: boilerplate and repetitive writing tasks can be largely offloaded to AI. Thomson Reuters notes that in-house lawyers using generative drafting tools significantly cut their writing time, allowing them to focus more on strategic tasks (The promise of AI-powered legal drafting for in-house teams). In consulting, we can expect similar outcomes: consultants will spend less time wordsmithing and more time on high-value analysis and client interaction. Notably, the Harvard/BCG study on consultants using GPT found not only were tasks completed **25% faster** and also **with 40% higher quality** (Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality - Working Paper - Faculty & Research - Harvard Business School) – indicating that AI-assisted writing can actually *improve* the final product, not just speed it up.

Current vs. Future-State: *Today,* writing a 30-page report means starting from a blank template, lots of copy-paste from prior docs, and multiple review cycles for tone and accuracy. *In the future-state with on-premises AI,* the consultant is more of an editor and strategist: the AI produces a solid draft populated with data and references, and the consultant fine-tunes the narrative and ensures it meets the client's needs. Turnaround for delivering a draft to a client could go from two weeks to a few days. For the C-suite, this means the firm can **serve more clients or dedicate more attention to client relations** instead of internal documentation work. It's a direct productivity boost that translates to either higher throughput or reduced labor costs (or a mix of both).

Conclusion: Investment vs. Value

Across these functions, the ROI for implementing an on-premises AI is strongly positive and compelling.

The **productivity gains** and time savings translate to concrete financial value that far outweighs the cost of the system. There are also significant strategic benefits beyond the dollar value, such as client satisfaction, employee morale, time to market, and more.

Let's summarize the impact:

- **Financial ROI:** If an on-premises AI implementation (hardware, software, deployment, training, and maintenance) costs, say, $120k in the first year (let's assume a generous budget for a high-quality solution), the estimated **first-year benefits in accelerated productivity (9522 hours, $714,000) already yield a ~6x ROI (600% return)**. In subsequent years, costs would drop (one-time setup costs paid) while benefits grow as usage increases and the model is further tuned, leading to an even higher ROI (possibly **400%-600% in following years**). This bottoms up evaluation aligns with global trends: an IDC study found that on average companies are seeing **$3.5 in return for every $1 invested in AI**, and **top performers see up to $8 in ROI** (Measuring ROI of Generative AI Adoption | by Dr. Gopala Krishna Behara | Medium).

- **Intangible and Strategic Benefits:** Beyond the raw hours and dollars, the on-premises AI brings **qualitative advantages**:
  - **Faster Project Delivery:** When research, writing, and analysis tasks speed up, project timelines shrink. Delivering results faster can be a competitive differentiator that helps win business. It also allows the firm to handle more projects concurrently, directly boosting revenue potential.
  - **Higher Quality & Consistency:** The AI acts as a guardian of consistency – ensuring that advice given to clients is based on the latest information and that documents maintain a high standard. Fewer errors or omissions mean less firefighting and risk mitigation down the line. As noted in the BCG study, AI augmentation led to **40% higher quality outputs** on knowledge (Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality - Working Paper - Faculty & Research - Harvard Business School). Quality improvements strengthen client trust and can lead to repeat business.
  - **Employee Satisfaction & Capacity for Innovation:** By offloading drudge work, consultants can focus on more fulfilling aspects of their job (creative problem solving, client interaction, innovative thinking). This improves morale and can reduce burnout/turnover – an indirect cost saving on recruitment. It also means

the team can devote time to developing new service offerings (e.g., perhaps using the on-premises AI to analyze client data for insights, a new value-add service) – driving growth.

- **Data Security & Compliance:** The choice of an **on-premises** solution is crucial for a pharma consulting firm dealing with sensitive IP. All the aforementioned benefits come **without compromising confidentiality**, since the on-premises AI and vector database reside behind the company firewall. **This avoids the regulatory and legal risks that would come with sending proprietary data to a third-party cloud**. In fact, **this could be a selling point to clients** – that the firm uses cutting-edge AI internally in a secure manner to serve clients better, differentiating from less tech-savvy CROs and even those leveraging other AI solutions without this privacy focus.

- **Scalability of Expertise:** The on-premises AI effectively **bottles the firm's collective expertise and makes it available on-demand**. This makes the firm more resilient – for example, if a senior expert is on leave, the junior staff can still retrieve that expert's knowledge via the AI. It lessens key-person risk and ensures continuity of service at a high level.

From an investment standpoint, C-suite stakeholders will want to know the payback period and ongoing value. In this case, the **payback period for the AI investment is likely well under one year**.

Every dollar spent on the AI yields multiple dollars in returns via cost savings or additional revenue capacity. Moreover, the risk of the investment is mitigated by choosing use cases with proven value (as we did with the functions above). Many of these gains are backed by industry benchmarks and case studies, giving confidence that they are *realistic* – e.g., **knowledge workers have demonstrably completed tasks ~25% faster with on-premises AI assistance (Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality - Working Paper - Faculty & Research - Harvard Business School)**, and **drafting documents is routinely 50% quicker with (The promise of AI-powered legal drafting for in-house teams)**. This is not hype or science fiction; it's happening now in forward-looking organizations.

In summary, deploying an on-premises AI in a veterinary pharma consulting firm is a **high-ROI proposition**. It directly addresses pain points in the firm's workflow, turning time sinks into opportunities for value creation. The firm can expect increased productivity (potentially on the order of 20+% per consultant), improved work quality, and the ability to scale its services without linear growth in headcount. Financially, the investment would likely pay for itself within the first year, and then continue to deliver returns in the form of labor efficiency and enhanced revenue for years to come. For the C-suite evaluating this, the recommendation is: **embracing a tailored AI solution will not only save money but also sharpen the firm's competitive edge in the fast-evolving consulting landscape**. In an industry where knowledge and speed are key, failing to leverage AI is arguably a bigger risk than the investment itself. Thus, the ROI here is not just in dollars, but in ensuring the firm remains at the forefront of innovation, delivering superior value to its clients.

Sources:

- Productivity gains from AI in knowledge work (BCG/HBS (Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality - Working Paper - Faculty & Research - Harvard Business School).
- Thomson Reuters on drafting efficiency (The promise of AI-powered legal drafting for in-house teams).
- Time spent on information search (McKinsey/IDC (Various Survey Statistics: Workers Spend Too Much Time Searching for Information - Cottrill Research).
- Morgan Stanley case on internal GPT ad (Shaping the future of financial services | OpenAI) (Shaping the future of financial services | OpenAI).
- Regulatory intelligence automation (DIA Global (Large Language Models: Extracting and Summarizing Regulatory Intelligence from Health Authority Guidance Documents).
- Literature review time reduction (Elicit AI ex (Elicit: The AI Research Assistant).
- Generative AI in clinical trial design (Simplifying Clinical Trial Protocol with Generative AI).
- Market analysis with AI (How we generative ai impact business performance.).
- ROI benchmarks (IDC, etc.) for AI investors (Measuring ROI of Generative AI Adoption | by Dr. Gopala Krishna Behara | Medium).