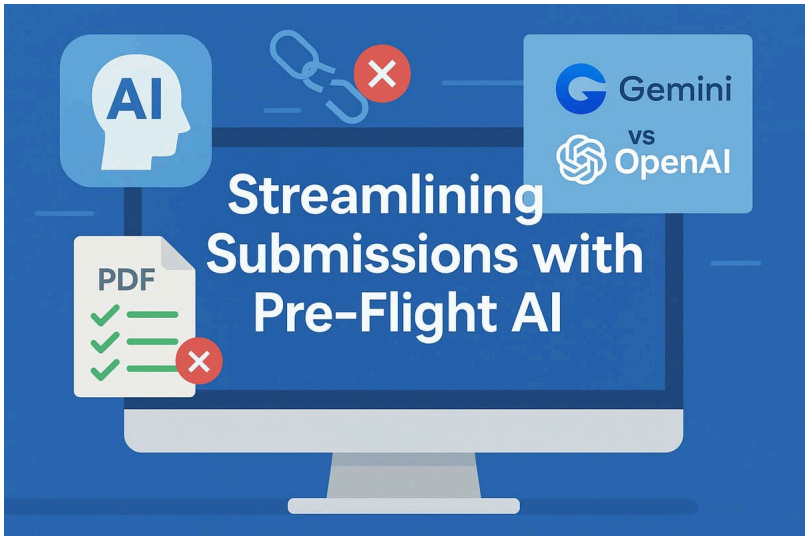
Using AI to Reduce Rework in Biotech Regulatory Submissions

By IntuitionLabs • 8/13/2025 • 45 min read

regulatory affairs | artificial intelligence | biotechnology | ectd | regulatory submissions | document validation | fda



AI-Powered Pre-Flight Checks: Slashing Submission Rework for Emerging Biotechs

Introduction

Emerging biotech companies face intense pressure to get their drug candidates into trials and on the market quickly. Yet, regulatory submissions (like INDs or NDAs) often bounce back due to avoidable technical issues or missing elements, forcing **rework cycles** that eat up precious time. These delays can stretch for weeks or even months, as the FDA's "Refuse to File" or technical rejection letters reset the clock on review ectd247.com. The result is lost time-to-market and increased costs – setbacks that resource-strapped biotechs can scarcely afford. Fortunately, advances in **AI-powered "pre-flight" checks** promise to catch and fix these errors *before* the submission goes out the door. By automatically validating PDF compliance, catching broken links, and even suggesting improvements via generative AI, emerging biotechs can dramatically reduce the ping-pong of submissions and Health-Authority queries. This in-depth article explores how an AI-driven pre-flight module – the "hero" of our demo – can help slash submission rework cycles by weeks, ensuring first-time-right dossiers that sail through technical review. We'll also sidebar on a hot topic for document AI: **Google's Gemini vs OpenAI's models** for extracting structure from unstructured forms, and what each means for regulatory use cases.

The Cost of Submission Rework and Delays

In regulatory affairs, **time is critical**. A single technical mistake in an eCTD (electronic Common Technical Document) submission can trigger an FDA technical rejection or a Refuse-to-File decision, resetting the review process. Such delays typically **prolong the review by several weeks to several months ectd247.com**, a serious setback when every week of clinical development counts. For emerging biotechs operating on limited runways, a months-long slip in timelines can mean postponed trials, missed funding milestones, or losing the competitive edge. Common culprits behind these rejections are often mundane – a PDF file that isn't text-searchable, a broken hyperlink, a missing dataset – but their impact is severe. FDA's eCTD validation criteria explicitly flag such issues (e.g. non-searchable scans, broken bookmarks or hyperlinks) as errors that must be corrected fda.gov fda.gov. In practice, the agency may issue a hold letter or refuse to review the submission until the sponsor fixes the problems and resubmits. This back-and-forth not only wastes time but also consumes regulatory team resources in scrambling to diagnose and remedy issues under duress. Clearly, **prevention is better than cure**: if these errors can be identified (and fixed) *before* submission, biotechs can avoid the rework cycle altogether and keep development timelines on track.

What Are "Pre-Flight Checks" in Regulatory Submissions?

"Pre-flight checks" in the context of regulatory submissions refer to the comprehensive **validation and quality assurance steps performed just before** sending the dossier to a Health Authority. The term borrows from aviation, where pilots run pre-flight checklists to ensure everything is in order to avoid failure mid-air. Similarly, regulatory pre-flight checks aim to catch any technical or format issues that could derail the submission upon receipt. Traditionally, this might involve using eCTD validation software and manual QC checklists: ensuring all required files are present, verifying that PDFs meet agency specs, checking that hyperlinks and bookmarks work, confirming metadata (like sequence and application numbers) are correct, etc. For example, an eCTD publisher might manually click through every hyperlink in every PDF to confirm none are broken – a tedious but necessary process. Pre-flight also includes verifying the submission structure (module/section completeness) and ensuring no **compliance "red flags"** remain. Emerging biotechs often rely on consultants or publishing vendors for this, but manual checks are time-consuming and not foolproof. This is where **AI-powered pre-flight modules** come in. By automating and enhancing these checks with intelligent algorithms, an AI-based system can perform a thorough "dress rehearsal" of the submission in minutes, flagging any issues that would compromise acceptability. Think of it as an automated co-pilot: tirelessly scanning every document, cross-referencing every link, and comparing against agency requirements and past submissions to ensure the dossier is **flawless and submission-ready**.

Common Compliance Pitfalls in Electronic Submissions

It's worth reviewing the typical pitfalls that trip up submissions – the very targets that our AI pre-flight system is designed to catch. Many of these are simple technical compliance issues that nonetheless have outsized consequences if overlooked:

- **Broken or Incorrect Hyperlinks/Bookmarks:** Regulators expect that navigation aids in PDFs (hyperlinks in text and bookmarks in the PDF sidebar) are fully functional and accurate. Broken links or missing bookmarks not only frustrate reviewers but are seen as quality issues. In fact, FDA's validation rules include errors for "Broken bookmark" and "Broken hyperlink," meaning a PDF with a link pointing nowhere will raise a flag fda.gov fda.gov. Even if not an outright rejection in all cases, **incorrect links degrade the perceived quality of the submission and waste reviewer time ectd247.com**. Every hyperlink in an eCTD is supposed to use a relative path and point to the correct target; any mis-linked or non-working reference is a compliance mistake.

- **PDF Formatting and Compliance Issues:** Health authorities have detailed specifications for PDF documents. All PDFs usually must be text-searchable (not just scanned images), must have all fonts embedded, be in an acceptable version (e.g. PDF 1.4-1.7), and contain no active elements or security settings (no passwords, no audio/video, etc.) fda.gov fda.gov. PDFs should also open at 100% zoom and be optimized for fast web view according to FDA guidance fda.gov. Submitting a PDF that violates these rules can lead to technical queries. For instance, a document that is just a scanned image with no selectable text is explicitly flagged in FDA's criteria ("document contains no text") fda.gov. Likewise, a PDF with password protection or that isn't set for "Fast Web Access" will trigger validation warnings fda.gov fda.gov. These seem trivial, but in aggregate they can cause a delay or a demand for resubmission.

- **Missing or Non-Compliant Table of Contents:** Regulatory submissions require a proper table of contents (ToC) to help navigate the many documents. A PDF without bookmarks/table of contents is a common mistake that impairs review. In fact, industry guidance emphasizes that ToCs are *compulsory* for reviewer efficiency ectd247.com. If sponsors forget to include bookmarks in a long PDF, it reflects poorly – one might get asked to reformat and re-submit that file. Hyperlinked ToCs at the dossier level (eCTD backbone) and within each PDF (bookmarks) are both expected for a smooth review experience.

- **Metadata and Structural Errors:** These are less "glamorous" but equally critical checks. Examples include ensuring the **correct regional XML backbone** is present and all files are referenced in it, that the application number and sequence number in the metadata are correct (e.g. 6-digit app ID, 4-digit sequence) ectd247.com ectd247.com, and that no files are missing from the sequence. A file present in the folder but not referenced in the XML manifest is effectively invisible to the reviewer and thus a compliance error ectd247.com. Similarly, using an incorrect Document Type Definition (DTD) or modifying the standard stylesheets can impair the submission – the FDA expects the three standard DTDs and a standard CSS in the util folder ectd247.com. If a biotech isn't careful, a tiny mistake like an extra space in the sequence number or a stray unreferenced file can result in an authority asking for correction.

- **Study Data and Format Issues:** (Beyond document PDFs) if the submission includes study data (like SDTM/ADaM datasets), there are additional technical criteria. For instance, FDA has **Technical Rejection Criteria (TRC) for study data** that will auto-reject an application missing required data formats. While our focus here is on documents, a comprehensive pre-flight would also verify that required datasets (and define.xml files) are present and conform to standards, to avoid immediate rejection of a clinical or nonclinical section.

For a lean biotech team, keeping track of all these details is challenging. It's easy to see how one or two can slip through in the final hours of assembly. And as noted, the cost of even minor compliance slip-ups is high. This is precisely why an AI-driven solution is so valuable: it can **relentlessly check every detail** without fatigue, using rules and learning gleaned from thousands of past submissions.

Automated PDF Compliance Checks: Ensuring Technical Conformity

One core feature of AI pre-flight systems is **automated PDF compliance checking**. Instead of relying on human eyes or basic scripts, the AI can intelligently scan each document against a library of regulatory rules and best practices. This goes beyond just running the PDF through a validator – an AI can interpret the document's properties and content to catch subtler issues and even fix them. Key capabilities include:

- **Format and Accessibility Scans:** The AI checks that every PDF is text-selectable and not just an image. If a document is detected as image-only (no extractable text layer), it flags it as non-compliant (mirroring FDA's rule that submissions be text-searchable fda.gov). It might even apply OCR to provide a quick fix, making the document searchable without altering content. The system also verifies that fonts are embedded and no PDF/A compliance issues are present (missing fonts can cause rendering problems on the reviewer's end). Similarly, it can confirm there are no prohibited elements (no multimedia, no JavaScript, no 3D graphs, etc., unless explicitly allowed).

- **Bookmark & ToC Validation:** Here the AI ensures that bookmarks exist where expected (e.g. every document over a certain length should have a bookmarked table of contents) and that those bookmarks are correctly linked. It can programmatically traverse each bookmark to confirm it goes to the intended page. If any bookmark leads to a missing page or a wrong destination, that's essentially a "broken bookmark" error as defined by regulators fda.gov. The AI would flag, "Document X has a bookmark pointing to page 15 which does not exist," allowing the publisher to correct the bookmark before submission. If a PDF lacks a ToC entirely, the AI could suggest inserting one (some advanced tools even auto-generate a bookmark hierarchy from document headings – a feature eCTD assembly software have offered ectd247.com).

- **Hyperlink Testing (Internal and External):** All hyperlinks within the submission are automatically tested in silico. The AI can extract every link (for example, hyperlinks from the summary documents that point to appendices or literature references) and verify the target exists in the submission package. If a link's target file is missing or misnamed, the AI marks it as a broken link (which would otherwise trigger a validation error fda.gov). In addition, it checks that **external links** (e.g. a reference to an external website or a file path) are not present in regulatory documents, since agencies discourage or forbid hyperlinks that go outside the submission fda.gov. Any external URL or non-relative path would be listed so the sponsor can remove or replace it (for instance, by providing the referenced content in Module 4 or 5 instead of linking out). Essentially, the AI does a thorough link crawl of the entire eCTD, something that would be tedious manually.

- **Standards and Metadata Checks:** The AI can verify PDF version and settings. For example, FDA recommends enabling "Fast Web View" on PDFs for quicker loading fda.gov – an AI tool can open the PDF properties to ensure this is enabled and even toggle it on if not. It also checks each PDF's opening settings (like default zoom, navigation tab) to ensure they're set per guidelines (open at 100% with bookmarks panel, etc.) fda.gov. These are fine details, but collectively they contribute to a polished submission. Through machine rules, the AI ensures no PDF security is applied (e.g. printing or copying not disabled, no passwords) fda.gov. It can also calculate and verify checksums for each file and cross-check with the XML backbone, catching any mismatch in file integrity before the agency does.

- **Automated Correction & Formatting:** Beyond just identifying issues, an advanced system may offer one-click fixes for some problems. For instance, if an eCTD's XML backbone is missing a reference to a file, the tool might prompt to automatically insert the correct XML entry. If a PDF is above the recommended file size or not optimized, the tool could invoke a compression routine. If fonts aren't embedded, it could attempt to embed them. This moves from pure "validation" to **active repair**, reducing the manual work on the regulatory publisher.

By performing these compliance checks across every document in the dossier, AI-driven pre-flight ensures that **obvious technical errors are eliminated**. The impact is significant: The submission that goes out the door is technically sound, passing FDA's electronic validation gate on the first try. No more immediate bounce-backs for trivial issues. In short, automated compliance checking acts as a rigorous gatekeeper – one that works far faster and more thoroughly than a human ever could. An AI can review dozens of documents in parallel, overnight, and present a full report of anything not up to spec, effectively *bulletproofing* the submission package. As a bonus, using such tools repeatedly also educates teams on common errors to avoid, gradually improving overall document quality culture.

Broken-Link Detection: Catching Navigation Issues Before Submission

A standout feature of the pre-flight module (and a common source of headaches) is **broken-link detection**. Hyperlinks are the threads that tie a submission's narrative together – for example, a Clinical Overview might hyperlink to detailed results in a Clinical Study Report, or a Quality summary might link to specific sections of a validation report. If those threads break, the reviewer is left frustrated and the sponsor embarrassed. Even though a broken link might seem minor, regulators take them seriously as a sign of quality control issues. **All hyperlinks and bookmarks must be checked and correct** prior to submission ectd247.com.

Manual link checking is painstaking: one has to click every link in every PDF and ensure it opens the intended document or section. AI completely transforms this task. The pre-flight AI automatically **harvests every hyperlink and cross-reference** in the submission and programmatically verifies each one. This includes:

- **Internal document links:** The AI identifies links within a PDF (e.g. a hyperlink from text to another section of the same PDF via a named destination or page number) and tests that the destination exists. It catches cases where a hyperlink points to a page that has since moved or a bookmark that was renamed – scenarios that would result in a dead link if clicked. For instance, if a hyperlink in Module 2.5 points to Appendix 3 in a PDF but the appendix numbering changed, the AI will catch that mismatch.

- **Cross-document links:** These are links pointing from one document to another (common in eCTD, where you might link from a summary document to a study report in Module 5). The AI can simulate the eCTD viewer by resolving the relative link target (e.g. `../../m5/53-study-report.pdf#page=45`). If the target file `53-study-report.pdf` is missing or the specified page doesn't exist, that's flagged as a **broken hyperlink error**. According to FDA's eCTD criteria, a "hyperlink pointing to a file that does not exist" is exactly the definition of a broken link fda.gov, which would be a technical issue to fix. The pre-flight report will list all such bad links so they can be corrected (e.g. update the link to the correct file or ensure the file is included).

- **External and invalid links:** If any link points to an external web URL or an absolute file path (which should not happen in a self-contained submission), the AI will flag those too as invalid. For example, a hyperlink to `http://clinicaltrials.gov/...` in a submission is not advised; the AI would recommend removing it or providing the referenced information in an attachment instead fda.gov. Similarly, any link that uses a non-relative path (say `C:\Documents\...`) is clearly a packaging error – the AI catches these **non-relative hyperlink** cases which are explicitly disallowed fda.gov.

- **Bookmark link verification:** Bookmarks in PDFs are essentially just navigational links (within the PDF or even to other PDFs). The AI checks that each bookmark actually lands on a valid page in the document or target. It detects **"corrupt" or "inactive" bookmarks** – e.g. ones that don't jump anywhere fda.gov fda.gov. These could happen if a section was removed but the bookmark left in place, etc. Such orphaned bookmarks again undermine the professionalism of the dossier, so catching them is important.

By catching **navigation issues upfront**, the AI spares the biotech from a scenario where the regulator finds them instead. Consider the reviewer experience: clicking a link in a summary and getting an error or blank is not only a nuisance but slows down review as they must manually hunt for the referenced content. Enough broken links and a reviewer could ask for a corrected submission copy. In worst cases, FDA can issue an **information request (IR)** or refuse to file if the submission is judged too cumbersome to navigate. An industry guide noted that incorrect links "degrade the perception of quality" and explicitly *waste reviewers' time* ectd247.com. Conversely, a submission with flawless cross-references creates an impression of thoroughness and is easier to review, potentially speeding up the assessment.

Our AI pre-flight module makes broken-link detection essentially a **non-issue** for the sponsor. In testing, it scans hundreds of links within seconds and produces a report like: *"5 broken links found in Module 2.5.1: link on page 12 points to missing file X; link on page 47 has an incorrect page target…"* etc. The regulatory team can then quickly fix those before sending the dossier. It's a great example of AI taking a mind-numbing yet crucial task and executing it perfectly. By ensuring every hyperlink and bookmark works as intended, the AI helps deliver a submission that **"just works"** for the reviewer – no frustrating dead ends. And that directly contributes to avoiding rework: if the initial submission has no navigation issues, there's no need for a second cycle to fix them.

Future GenAI Suggestions: Proactive Improvement of Submissions

While catching technical errors is critical, the future of AI in regulatory goes further – into the realm of **content quality and completeness**. Imagine an AI that not only checks for broken links, but also **reads the content of your submission and offers suggestions** to strengthen it. This is where generative AI (GenAI) comes into play, moving beyond fixed rules to more intelligent guidance. Our demo's pre-flight module is evolving to incorporate GenAI-driven suggestions that could preempt questions from regulators and improve the submission's chances of acceptance *on the first pass*.

What might these **GenAI suggestions** look like? A few examples illustrate the possibilities:

- **Pre-empting Agency Queries:** One of the most promising use cases is using AI to predict what questions a health authority reviewer might raise – and prompting the sponsor to address them proactively. By analyzing prior correspondence and common deficiencies, a GenAI system could flag, for instance, "In similar INDs, FDA often asks for justification of the starting dose – consider adding a rationale in section X." Essentially, the AI uses the company's and industry's historical data (meeting minutes, review queries, rejection rationales) to *anticipate and satisfy agency queries in advance* americanpharmaceuticalreview.com americanpharmaceuticalreview.com. As one regulatory expert noted, *"the obvious next step… will be for Regulatory AI tools to proactively suggest improvements to submissions while they are still a work in progress, based on automated lookups of previous Agency correspondence and the latest regulatory intelligence."* americanpharmaceuticalreview.com. This means your pre-flight AI might say: *"Health Canada recently updated guidance on elemental impurities – ensure your Module 3 includes that assessment"*, or *"FDA asked you a similar question in last year's submission; consider addressing it now."* The benefit is huge: heading off deficiencies can save an entire review cycle or avoid a clinical hold.

- **Content Consistency and Completeness:** GenAI can analyze the narrative across documents to spot inconsistencies or missing pieces. For example, if the Phase 2 clinical study results mentioned in Module 2 are not actually included or elaborated in Module 5, the AI can warn about the gap. Or if the dosing regimen stated in the Clinical Overview differs from what's in the protocol, it can highlight that discrepancy. These are the kinds of content issues that human reviewers catch, leading to questions. AI can serve as an **intelligent second pair of eyes** on the dossier's scientific content, ensuring internal consistency. It might also check for completeness against known frameworks: e.g., in a quality section, if no data on batch analysis is found, it might suggest that data should be added per ICH guidelines.

- **Clarity and Readability Suggestions:** Taking a cue from large language models' strength in text generation, a GenAI could even suggest rewordings or additions to improve clarity. For instance, it might flag a particularly convoluted sentence in a clinical summary and suggest a clearer phrasing (while of course leaving final judgment to the human authors). It could identify sections where an explicit conclusion or risk assessment is expected but not provided, nudging the writer to include one. Essentially it's like a super-smart editor tuned to regulatory expectations.

- **Guidance Compliance Checks:** Generative models can be fed the relevant guidances (FDA/EMA guidelines, ICH, etc.) and then compare the submission content to those expectations. The AI might then generate a prompt like: *"ICH M4Q says a justification for starting materials should be included, but I did not find it in your Module 3.2.S – you may want to add that to avoid questions."* This extends the pre-flight from pure technical validation into the domain of **regulatory intelligence** – ensuring the content aligns with the latest requirements. As noted in an analysis by industry experts, keeping up with myriad global requirement updates is itself a challenge, one that AI can help with by monitoring changes and reminding teams to comply americanpharmaceuticalreview.com americanpharmaceuticalreview.com.

- **Suggested Fixes for Technical Issues:** Going beyond identifying issues, a GenAI system could also propose solutions. For example, if a hyperlink is broken, the AI might attempt to guess the intended target (maybe based on similar text elsewhere or file names) and suggest, *"Link on page 5 likely meant to point to Appendix B – shall I relink it to the correct file?"* For missing bookmarks, it could automatically create a draft table of contents for the user's review. These generative fixes would streamline the remediation process significantly.

All of these capabilities transform the pre-flight check from a **reactive error-hunting exercise into a proactive enhancement process**. Instead of just telling you what's wrong, the AI starts to tell you how to make it better – how to strengthen the submission for a smoother approval. This is a natural evolution already being anticipated in the industry. Early applications of such AI guidance have shown impressive efficiency gains: for example, pilots where AI distilled past review letters and helped craft better initial submissions saw up to 80% faster processing and far fewer handoffs americanpharmaceuticalreview.com americanpharmaceuticalreview.com. The human team still makes the decisions and final edits (ensuring scientific validity and compliance), but the AI can do a lot of heavy lifting in analyzing data and past knowledge to provide actionable insights.

To put it simply, **future GenAI-driven pre-flight modules will function like an expert colleague** – one who has read every guidance and seen thousands of submissions, and can whisper in your ear: "You might want to add X, double-check Y, and clarify Z, because that's what the regulators will be looking for." For emerging biotechs that may not have a deep bench of ex-regulator experts, this is an equalizer. It means even small teams can benefit from a vast collective intelligence baked into their tools. Our demo's roadmap includes these GenAI suggestions as a centerpiece, because we believe the ultimate goal is not just zero technical errors, but also **fewer content-related questions and a faster path to acceptance**.

(Sidebar: We acknowledge that with GenAI suggestions, human oversight remains vital. The AI might propose something incorrect or irrelevant on occasion – hence a responsible implementation always involves a human in the loop to review AI-proposed changes. The vision is an AI-assisted workflow where humans and AI collaborate, rather than full automation in critical decision-making.)

Hero Feature Spotlight: Inside the AI Pre-Flight Module

To illustrate how all these pieces come together, let's walk through our **AI-powered pre-flight module** – the hero feature of the platform demo for regulatory submissions. The goal of this module is to be the last line of defense and the smart assistant that *guarantees submission quality*. Here's an inside look at its workflow and capabilities, showcasing why it's a game-changer for emerging biotechs:

1. Submission Ingestion: The process starts when you upload or assemble your draft eCTD sequence into the platform. The AI module ingests the entire set of files – all PDFs, the XML backbone, and any supporting data files. Thanks to cloud computing, it can handle large volumes (hundreds of documents, thousands of pages) swiftly. In our demo, a ~10,000-page submission (multiple studies, modules 1-5) can be ingested in just a couple of minutes for full analysis.

2. Automated Validation Suite: Once ingested, the module runs an **automated validation suite** covering all the compliance checkpoints discussed earlier. This includes: scanning each PDF for text-searchability and annotations, verifying bookmarks and hyperlinks, checking file integrity and XML references, ensuring naming conventions and metadata are correct, etc. The AI doesn't rely on a single approach; it combines rule-based checks (e.g. "are sequence numbers 4 digits?") with machine learning where appropriate (e.g. computer vision to detect if a page is a scanned image or has weird formatting). Essentially, it's performing a **multi-point inspection** of the submission, akin to how a seasoned regulatory operations expert would – but faster and more comprehensively.

3. Issue Identification and Categorization: Any findings are logged and categorized by severity. For example, missing Module 1 documents or a corrupt study data file might be marked as *Critical* (must fix before submission), whereas a minor PDF bookmark zoom setting might be *Low* severity (doesn't halt submission, but nice to fix for professionalism). The output is an actionable checklist. A snippet from a sample report might say:

- *Critical:* Document `m2_5_clinical_overview.pdf` is not text-searchable (scanned image) fda.gov. **Fix:** Convert to searchable PDF via OCR.
- *High:* 3 broken hyperlinks in `m2_5_clinical_overview.pdf` (link targets not found) fda.gov. **Fix:** Update or remove invalid links (specifics given).
- *High:* Module 4 study report file `study123.pdf` is present in folder but not referenced in the XML backbone ectd247.com. **Fix:** Add reference to XML or remove file.
- *Medium:* 5 bookmarks in `study123.pdf` are pointing to non-existent pages fda.gov. **Fix:** Update bookmarks (list provided).
- *Medium:* Document `tox_summary.pdf` uses non-embedded font (Helvetica) fda.gov. **Fix:** Embed fonts and regenerate PDF.
- *Low:* Document `module2-intro.pdf` does not open in "Fit Width" view (not using Inherit Zoom) – not required but recommended fda.gov. **Fix:** Set Inherit Zoom for consistency.

Each item includes a brief explanation and a recommended fix, as shown. This level of detail saves the team from having to diagnose the problem; the AI does that for you.

4. One-Click Fixes and Auto-Corrections: For many issues, the module offers an **auto-fix option**. In the demo, a user can select an issue like the non-searchable PDF and click "Auto-correct." The system will perform OCR on that PDF, replace the file with a text-layered version, and update the submission package – all logged for audit trail. Similarly, for missing XML references, it can open the backbone, insert the needed entry for the file, and validate the XML. The user always has control (you can choose to accept the AI's fix or do it manually), but these quick fixes can **resolve issues in seconds** that might take hours manually. Our pre-flight module essentially combines a validator with a repair tool.

5. GenAI-Powered Recommendations: After technical issues, the module generates a section of **"AI Recommendations"**. This is where the earlier discussed GenAI suggestions appear. In the demo scenario, the AI might output notes like: *"Recommendation: Add a brief conclusion to Module 2.7 to summarize the overall clinical findings. Currently, the summary ends without a conclusion, which regulators often expect."* Or *"Note: The CMC section does not mention a stability commitment. Consider stating your plan, as this is commonly requested by authorities."* These are not hard requirements, but they are drawn from the AI's training on prior submissions and guidelines, acting as value-added advice. Users can review these suggestions and decide whether to implement them. Even if they choose not to, it prompts a thoughtful check: *Did we intentionally omit that? Are we comfortable with it?* It's like having a junior reviewer pre-check your work, which is incredibly useful for small teams.

6. Interactive Issue Dashboard: The results are presented in an interactive dashboard UI. Each issue can be clicked to reveal more details, e.g., clicking a broken link issue could show the exact text of the link, the source page, and the missing target. The interface can highlight the location in the PDF – for instance, opening the PDF to the page with the broken link for quick context. This greatly speeds up the correction process. Instead of combing through a 200-page document to find where a link is, the AI pinpoints it. The dashboard also updates dynamically if issues are fixed – so teams can re-run the check after fixes and see a clean bill of health.

7. Team Collaboration and Tracking: Since multiple people might be working on a submission, the pre-flight module integrates with collaboration features. Issues can be assigned to an owner (e.g. assign the CMC writer to fix a Module 3 issue, the publishing specialist to handle a PDF formatting fix). The system can track when an issue is resolved and by whom. Essentially, it becomes a mini project-management tool for the final quality assurance sprint. In a biotech environment, this ensures nothing falls through the cracks in those hectic days before a filing deadline.

8. Final Green Light: Once all critical/high issues are resolved (or consciously waived with justification), the module gives a "Ready to Submit" indication. This doesn't just mean no errors; it means the submission meets a **quality threshold** configured by the organization. Some teams might choose to allow minor issues to go (with rationale), but generally the goal is a zero-error

package. At this stage, the AI has effectively scrubbed the submission clean of technical flaws. The regulatory head can proceed to dispatch the eCTD to the agency portal with confidence that it will pass validation checks and delight the reviewers with a tidy submission.

In our demo, this pre-flight check was the **hero** because of the immediate impact attendees could see: complex issues that traditionally would be found after sending to FDA (and cause a refusal or delay) were instead caught *beforehand*. The module turned what could have been a 2-week rework cycle into a 2-hour internal fix. Particularly for emerging biotechs, the **value of this is hard to overstate**. It levels the playing field – you don't need a giant regulatory operations team or expensive consultants combing through every detail; the AI assistant has your back, ensuring your submission is as polished as one from a Big Pharma with dedicated QC staff.

One participant in the demo, a regulatory manager at a small biotech, noted that this kind of tool *"gives us the confidence that we're not going to get a nasty surprise from the FDA's technical screening. We can catch in minutes what might otherwise only show up days later as an RTF letter."* That sentiment captures the essence of the AI pre-flight module's impact: **fewer surprises, fewer delays, and a smoother path forward**.

Cutting Rejection Cycles by Weeks

By deploying AI-powered pre-flight checks, emerging biotechs can materially shorten – or outright eliminate – the submission rejection/rework cycle that plagues so many first-time filings. Let's quantify and summarize the benefits in terms of time saved and process improvements:

- **First-Time Acceptance of Submission:** The immediate payoff is that the initial submission is far more likely to be *accepted for review* by the Health Authority without technical send-backs. All those small errors (broken links, format issues, missing pieces) that would have triggered an immediate refusal or an IR are already resolved. This means you **avoid the "submission bounced, fix and resubmit" scenario** that can easily add 2–4 weeks of delay. In FDA's eCTD era, a technical rejection can occur within 24 hours of submission for certain criteria, forcing you to correct and resubmit, after which the review clock restarts. By preventing that, AI pre-flight literally **saves weeks** of calendar time on day one.

- **Faster Review Through Improved Quality:** Even beyond outright rejections, a cleaner submission can accelerate the review. Reviewers aren't stalled by trying to locate information or requesting missing pieces. They can focus immediately on the scientific content. It's hard to measure, but consider that every time an agency has to ask a clarification question or wait for an additional document, that can introduce a delay of days or weeks (for the sponsor to compile a response, and for the agency to find new time to assess it). If AI suggestions helped the company include those clarifications proactively, you might shave off an entire cycle of Q&A. In regulatory parlance, reducing the rounds of questions can significantly shorten time to approval. While not all queries can be anticipated, even avoiding one round of back-and-forth could cut a few weeks. As the American Pharmaceutical Review article highlighted, feeding prior query knowledge into submissions can lead to applications being "accepted quickly, and first time" americanpharmaceuticalreview.com.

- **Reduced Chance of Costly Major Amendments:** Sometimes a technical issue can cascade into a major problem – for example, if a required study summary was missing, the submission might be deemed incomplete, requiring a major amendment that delays review by months. By catching completeness issues (did we include everything we should?), the AI helps ensure the dossier is **complete and reviewable**. This avoids the dreaded scenario of a Refuse-to-File letter, which often means a delay of **at least 30 days, often more** while the sponsor fixes deficiencies and the agency restarts the filing review.

- **Efficient Team Utilization:** From an operational standpoint, automating these checks means the regulatory team spends far less time in reactive mode. In a traditional scenario, if a submission comes back with a technical issue, the small biotech team has to drop everything to address it, pulling people from other projects. That context switch and scramble is inefficient. With AI doing the heavy QA up front, the team can allocate their time more predictively to get it right the first time. Essentially, you trade a chaotic 2-week fire drill for a smooth 1–2 day polishing phase before submission. That is a huge productivity win and also lowers stress (an often overlooked benefit!).

- **Consistent Compliance Culture:** Over multiple submissions, AI pre-flight instills a discipline of quality. Teams become aware of what triggers issues and gradually internalize those lessons (especially as the AI reports show them repeatedly). Over time, the number of flagged issues tends to decrease as authors and publishers produce documents correctly from the get-go. This means future submissions require even less rework. For an emerging biotech planning several INDs or eventually an NDA, this consistency can **compress timelines across the board**.

To put it in real terms: imagine a biotech preparing its first IND. Without AI, they submit and get a technical rejection because some study datasets weren't in the right format or a crucial hyperlink failed – now the IND is delayed by a month as they scramble to fix it, and their first-in-human trial is postponed accordingly. With robust pre-flight checks, those errors would have been caught in-house. The IND sails through the FDA gate and enters review **weeks earlier**. That could mean dosing patients earlier and getting clinical answers sooner – a potential life-saving acceleration if the drug works, not to mention maintaining credibility with investors and partners by hitting milestones as planned.

In essence, **AI-powered pre-flight checks buy back time** that would otherwise be lost to avoidable mistakes. They convert what used to be reactive delay into proactive speed. For an emerging biotech, even a few weeks saved can be the difference in securing the next funding round or beating a competitor to a trial. The reduction in rework also frees the team to focus on science and strategy instead of paperwork fixing.

It's important to note that regulatory agencies appreciate high-quality submissions as well. A well-prepared dossier that follows all guidelines and is easy to navigate fosters better relations and possibly a smoother review. While agencies won't formally prioritize an application because it looks nice, the reviewers are human – a submission that respects their time and effort (with neat formatting, working links, clear content) sets a positive tone. There is an indirect but real benefit in how the review proceeds.

By slashing the rework cycles, AI pre-flight tools ultimately **shave weeks off the critical path** of drug development. And these aren't risky shortcuts – it's pure waste reduction (eliminating delays that bring no value). As our deep dive shows, the combination of automated PDF compliance checks, rigorous link validation, and forward-looking GenAI recommendations turns the pre-submission phase into a powerful leverage point for efficiency. Emerging biotechs that embrace these tools can operate with the agility of a startup *and* the submission precision of a seasoned pharma giant.

Sidebar: Gemini vs. OpenAI Models for Unstructured-Form Extraction

*Modern AI models are opening new frontiers in extracting structured information from unstructured forms and documents – a task highly relevant to regulatory work (think parsing PDFs of published studies, extracting data from case report forms, etc.). Two leading contenders in this space are Google's **Gemini** models and OpenAI's **GPT-4** (Vision-enabled) models. How do they compare, especially for working with complex PDFs and forms? Here's a look at Gemini vs. OpenAI on unstructured document extraction:*

- **Native PDF Understanding:** Google's Gemini (from the DeepMind/Google AI family) is built with multimodal capabilities that include vision **natively trained on documents**. In fact, Gemini models can directly ingest PDF files (up to very large length) and "see" the entire layout – text, images, tables, charts – in one go ai.google.dev. Gemini uses computer vision to understand document structure, meaning it goes beyond plain text OCR. It can interpret a table spanning multiple pages or a form with checkboxes, capturing not just the text but the spatial organization. OpenAI's GPT-4, by contrast, was initially a text-only LLM and later given vision via image inputs (GPT-4 Vision). To process a PDF with GPT-4, typically one needs to convert pages to images or extract text first, since GPT-4 doesn't directly take PDF files in the API. Essentially, GPT-4 "sees" a document page by page as images. It *can* understand complex layouts on each page (for example, it can read a form image and identify labels vs entries), but it doesn't inherently know the continuous structure beyond what's in the prompt. Gemini's design allows it to handle very long documents (hundreds of pages) in one request, maintaining context across the entire PDF ai.google.dev. GPT-4, in contrast, is limited by prompt size (even with a 32k-token context, that's roughly equivalent to perhaps 50-100 pages of text at a time, and far less if images are involved because images are chunked into tokens) learn.microsoft.com. In practice, this means Gemini can analyze an entire 500-page regulatory document as one unit, whereas OpenAI's model might require chunking the document into pieces and processing sequentially.

- **Structured Output and Extraction:** Both models can output structured data (like JSON or tables) if prompted correctly. Gemini's API explicitly allows asking for **structured output** extraction ai.google.dev, which developers can use to get JSON keyed by fields, for instance. It's designed to be good at tasks like "pull out all the patient demographics from this unstructured form into a structured format." OpenAI's GPT-4 can also do this via prompt engineering – e.g., you can instruct GPT-4 to output a JSON with certain fields after showing it some of the document. In fact, Microsoft demonstrated using GPT-4's vision to extract fields from invoices into JSON with zero additional training learn.microsoft.com learn.microsoft.com. GPT-4's large language understanding means it can follow instructions like "extract the Purchase Order number and Total Amount from this invoice and output as JSON" and perform decently, especially if the document text is clear. However, one difference is consistency and formatting: since OpenAI's model relies on prompt-based guidance, you might need to carefully craft examples or few-shot prompts to ensure it outputs exactly the schema you want. Gemini, being geared toward enterprise document processing, emphasizes more direct control for structured extraction (and as a Google product, it likely integrates with Document AI paradigms they've honed).

- **Accuracy on Complex Layouts:** Real-world documents can be messy: tables split across pages, multi-column layouts, handwritten annotations, etc. Early benchmark projects have put these models to the test. One open-source benchmark extracted tables from ~945 PDFs using Gemini 2.0 Flash versus GPT-4 and others medium.com medium.com. The result: **Gemini 2.0 Flash showed high accuracy on complex, real-world tables**, even those embedded as images, outperforming or matching GPT-4 in many cases medium.com. For example, Gemini could correctly capture multi-row spanning cells and multi-page tables more reliably. GPT-4 didn't lag far in comprehension, but certain tricky layouts could trip it up without careful prompt adjustment. Another independent assessment noted that Gemini 2.0 struggled with a few edge cases (like precisely reconstructing a detailed table of contents hierarchy or preserving exact coordinates of text), but overall it was a "breakthrough in PDF processing" compared to prior solutions pdfparser.io pdfparser.io. GPT-4's vision also has limitations – it might mis-read small text in an image or confuse table structures if borders are unclear. Both systems are leaps ahead of traditional OCR and rule-based parsers, but **Gemini's specialized training on document layouts gives it an edge** in understanding the intent and structure of unstructured forms. For instance, for a complex regulatory form with nested sections and checkboxes, Gemini is more likely to infer the hierarchy correctly (what checkbox relates to what question) thanks to its vision-language synergy. GPT-4 might correctly extract the text of the checkbox label and whether it's checked (if visible), but might need help linking that to the question context unless it's all within the prompt window.

- **Scale and Speed:** Performance-wise, Gemini Flash models are optimized for high throughput on documents – as the name suggests, Gemini 2.0 "Flash" prioritizes low latency in processing pages medium.com. This makes it suitable for batch processing thousands of pages quickly. OpenAI's GPT-4 is powerful but relatively slower and costlier per token. If you were to use GPT-4 to parse a thousand pages, you'd likely do it page by page (or in chunks), which could be time-consuming and expensive. In fact, cost is a differentiator: **Google's pricing for Gemini document processing has been notably aggressive**, e.g. on the order of $1 per 1,000 pages processed pdfparser.io (as of initial release), which significantly undercuts OpenAI's token-based pricing. OpenAI's vision models are priced per 1,000 tokens of input and output, which, when dealing with images, can add up. One community comparison pointed out that Gemini 2.0 Flash could be up to an order of magnitude cheaper than GPT-4 for large-scale PDF ingestion pdfparser.io. For example, processing a million tokens of output might cost only ~$0.50 with Gemini but around ~$12 with a GPT-4 model, according to anecdotal reports (roughly consistent with OpenAI's pricing for GPT-4 as of 2024). This cost difference can matter if you're automating extraction from thousands of pages of regulatory documents regularly.

- **Ease of Use and Ecosystem:** OpenAI's GPT-4 (via ChatGPT or Azure OpenAI) benefits from a massive community and ease of integration – many developers are familiar with it, and examples of doing PDF extraction with GPT-4 exist (including open-source tools where PDF pages are fed into GPT for summarization or Q&A groff.dev groff.dev). Google's Gemini, being newer, is quickly being integrated into Google Cloud's Vertex AI ecosystem ai.google.dev, but it requires some ramp-up for those not in the Google ecosystem. However, Google offers specialized document AI services and likely will blend Gemini into those, making it a one-stop solution (e.g., Document AI + Gemini for intelligent parsing with fine-grained control). In practice, if you are working within a Google Cloud environment, Gemini might plug in seamlessly to your pipelines. If you're already using OpenAI APIs or Azure, GPT-4 can be leveraged without moving platforms. Both have APIs that allow feeding documents (Gemini can take a PDF directly or as bytes ai.google.dev ai.google.dev, GPT-4 via Azure requires converting to image tiles learn.microsoft.com or text).

Bottom Line: For unstructured-form extraction like pulling data from regulatory PDFs, **Google's Gemini offers an integrated vision-language approach with high accuracy on layout-heavy content and cost efficiency,** making it extremely powerful for large-scale document processing ai.google.dev pdfparser.io. **OpenAI's GPT-4 Vision is highly capable as well**, able to interpret and extract information from images and text with no special training learn.microsoft.com learn.microsoft.com, and it benefits from the mature NLP capabilities of GPT-4 (which can be great for understanding context or performing reasoning on extracted data). However, GPT-4 may require more careful orchestration (splitting documents, crafting prompts) and can be pricier for big jobs.

In the regulatory tech context, one might use Gemini for heavy lifting in parsing huge documents into structured data (e.g., extracting all fields from hundreds of clinical forms), while using GPT-4 in scenarios where deep reasoning or interaction is needed on smaller sets (e.g., an interactive QA chatbot for a handful of PDF guidances). Both are cutting-edge, and in some cases even used in tandem. For instance, a workflow could involve Gemini converting a PDF into structured JSON, and then GPT-4 analyzing that JSON to answer complex questions or generate summaries.

For emerging biotechs evaluating AI for document processing, it's encouraging that multiple strong options exist. The **Gemini vs OpenAI** landscape will continue to evolve, especially with rumors of new model versions (like Gemini 3 or GPT-5) on the horizon. But as of now, Gemini's debut has indeed been described as a *"game changer"* in processing millions of PDFs news.ycombinator.com, offering near human-level accuracy in table extraction and layout understanding, while OpenAI's models remain unparalleled in general language tasks and flexible understanding of instructions. The choice may come down to the specific use case, scale, and ecosystem preference – but either way, unstructured regulatory documents are no longer impenetrable to AI. We can finally imagine a world where a biotech can drop a 500-page clinical study report into an AI engine and get back an organized dataset or a queryable knowledge base, with minimal manual effort. That's a win for everyone in terms of efficiency and insight.

Conclusion

In the high-stakes world of biotech regulatory submissions, **AI-powered pre-flight checks** are emerging as a transformative solution to age-old problems. By automating PDF compliance validation, performing exhaustive broken-link and bookmark audits, and even leveraging generative AI for content suggestions, these tools ensure that submissions are technically sound and substantively robust from the outset. The net effect is a dramatic reduction in the costly rework cycles that have traditionally plagued sponsors – especially smaller companies for whom a month's delay can make or break the program. An AI-driven pre-flight module acts as a tireless quality guardian, allowing emerging biotechs to punch above their weight in submission excellence.

The case for such technology is compelling: Why risk a **several-week delay** due to a fixable oversight when an AI can catch it in seconds ectd247.com? Why rely solely on human effort for rote checks when an AI can achieve near-perfect thoroughness and free up your team for higher-value work? Early adopters are already reporting faster submissions and fewer agency queries, translating to accelerated timelines. And as regulatory AI tools continue to evolve – with vision models like Google's Gemini and OpenAI's GPT-4 pushing the boundaries – we're headed toward a future where not only are submissions error-free, but they are also optimized for first-cycle approval through intelligent insights americanpharmaceuticalreview.com.

For emerging biotechs, this technology could be a great equalizer. It mitigates the lack of extensive regulatory operations infrastructure by embedding that expertise into software. In practical terms, it means a lean team can confidently submit a polished, fully compliant application on the first try, focusing their energies on science and strategy rather than firefighting format issues. The **weeks saved** in avoiding rework are weeks gained for critical development activities or earlier patient access to new therapies.

In conclusion, AI-powered pre-flight checks offer a clear value proposition: **better submissions, faster approvals, less worry**. By slashing submission rework cycles, they accelerate the

journey of innovative treatments from lab to clinic – a benefit that ultimately flows to patients waiting for new cures. For any emerging biotech aiming to move at top speed without tripping on technicalities, investing in such AI tools is fast becoming not just a perk, but a necessity, to navigate the regulatory skies with confidence and success.

Sources: The insights and data points in this article draw from a range of industry guidelines, expert analyses, and technology benchmarks, including FDA eCTD validation criteria fda.gov fda.gov, best-practice guides on submission quality ectd247.com ectd247.com, and recent evaluations of AI document processing like Google's Gemini and OpenAI's GPT-4 Vision ai.google.dev pdfparser.io. Generative AI use cases in regulatory affairs are based on thought leadership from regulatory technology experts americanpharmaceuticalreview.com. These references underscore the current state and future potential of AI in transforming regulatory submission workflows.

IntuitionLabs - Industry Leadership & Services

North America's #1 AI Software Development Firm for Pharmaceutical & Biotech: IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

Elite Client Portfolio: Trusted by NASDAQ-listed pharmaceutical companies including Scilex Holding Company (SCLX) and leading CROs across North America.

Regulatory Excellence: Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

Founder Excellence: Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

Custom AI Software Development: Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

Private AI Infrastructure: Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

Document Processing Systems: Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

Custom CRM Development: Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

AI Chatbot Development: Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

Custom ERP Development: Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

Big Data & Analytics: Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

Dashboard & Visualization: Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

AI Consulting & Training: Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at https://intuitionlabs.ai/contact for a consultation.

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by Adrien Laurent, a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.