90-day AI agent pilot plan

Free from Atlas Bio (atlasbiosys.com). No form or sign-up. Version date: 2026-11-10.

How to use it

The plan (the 'Pilot plan' tab in Excel; the pages after this one in the PDF) has one row for each week-block in the table on the Atlas Bio answer 'How do you run a first AI agent pilot in a biotech?', plus one row for the separate gate before any GxP use that the page sets out. The weeks, focus, what gets done and written output are the page's own words (row 8: what gets done is from the page's Week 13 section; the page names no single written output for that gate). Fill in the four blank columns for your pilot: Owner, Date, Evidence (where the written output is kept) and Decision.

The page treats the week numbers as indicative; what matters is the order, and that each written output exists before the next step starts.

Where each row comes from

Each row quotes the page by section heading and gives the source the page cites for that step, as the page states it. Where the page cites NIST AI 100-1, the EMA reflection paper or the European Commission's draft Annex 22, the row also quotes the passage from that document; the page's own wording is a summary, the quoted passage is the source's text. Where the page names OWASP entries or FDA's draft guidance for a step, the row names that source and links it. Row 6 (evaluate) has no external source on the page.

EU GMP Annex 22 is a consultation draft (July 2025) and not in force. Its scope excludes generative AI and large language models, and it says such models should not be used in critical GMP applications. The page uses it as a good-practice habit for setting acceptance criteria in a discovery pilot (row 3), and to show why a separate gate is needed before any GxP use (row 8).

The inventory entry

Rows 3 and 7 create and update the agent's inventory entry. For the fields to record, use the Atlas Bio AI agent inventory register template (https://atlasbiosys.com/downloads/ai-agent-inventory-register-template.xlsx), explained at https://atlasbiosys.com/answers/what-should-an-ai-agent-inventory-in-pharma-rd-record.html. This plan does not repeat it.

What this plan does not do

It adds no thresholds, durations or results beyond the page: the week numbers are the page's indicative ones; choose your own success and stop criteria and record them before the supervised run. A pass at the go/no-go gate covers only the non-GxP use that was tested; any GxP use is a separate decision made through your quality system.

It is general information, not legal or regulatory advice. It does not validate any system. It is not published or endorsed by NIST, the European Medicines Agency, the European Commission or OWASP. You may copy, edit and use it inside your organization.

Page: https://atlasbiosys.com/answers/how-to-run-a-first-ai-agent-pilot-in-a-biotech.html

#	Weeks and focus	What gets done / written output (page)	Source (quoted)	Owner	Date	Evidence	Decision
1	**1-2** Choose the use case and owner	Shortlist tasks; pick one non-GxP discovery task where a wrong output affects only internal work; name the owner and the scientists who will review outputs; describe the task, inputs, outputs and what the agent must not do **Written output:** Intended use statement with the reason the task is non-GxP	Page, 'Weeks 1-2: pick a non-GxP discovery use case': “Write the intended use down before you build anything.” Page, 'Weeks 1-2: pick a non-GxP discovery use case': “Add one line on why the task is non-GxP and what would change that, because the classification needs revisiting if the output later moves towards regulated work.” Page, 'What a first pilot is for': “The European Medicines Agency's reflection paper on AI in the medicinal product lifecycle (September 2024) says the use of AI in drug discovery can be of low regulatory impact if non-optimal performance only affects the developer” Page, 'What a first pilot is for': “FDA's January 2025 draft guidance on AI to support regulatory decision-making for drugs and biologics states that it does not address AI used in drug discovery or for operational efficiencies that do not affect patient safety, drug quality or the reliability of study results.” NIST AI 100-1 (AI RMF 1.0), MAP 1.1: “Intended purposes, potentially beneficial uses, context-specific laws, norms and expectations, and prospective settings in which the AI system will be deployed are understood and documented.” NIST AI 100-1 (AI RMF 1.0), MAP 3.3: “Targeted application scope is specified and documented based on the system’s capability, established context, and AI system categorization.” EMA reflection paper on AI in the medicinal product lifecycle (September 2024), section 2.3.1 (Drug discovery): “application of AI/ML in the process of drug discovery can be low regulatory impact if non-optimal performance only affects the developer.” FDA draft guidance on AI to support regulatory decision-making for drugs and biologics (January 2025) (named by the page) https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle_en.pdf https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological				
2	**3** Scope the data	List the data the agent may read; choose public, synthetic or copied data; review the model provider's terms on data retention and use with IT security and legal; exclude personal data unless your privacy or legal function has reviewed its use **Written output:** Data scope note; provider terms reviewed	Page, 'Weeks 3-6: scope the data, then build in a sandbox': “Decide what the agent may read before it reads anything.” Page, 'Weeks 3-6: scope the data, then build in a sandbox': “do not let outputs made from copies of GxP records flow back into those records.” Page, 'Weeks 3-6: scope the data, then build in a sandbox': “OWASP's Top 10 for LLM Applications 2025 lists sensitive information disclosure (LLM02) among the main risks for LLM applications” OWASP LLM02:2025 Sensitive Information Disclosure (named by the page) https://genai.owasp.org/llmrisk/llm022025-sensitive-information-disclosure/				

#	Weeks and focus	What gets done / written output (page)	Source (quoted)	Owner	Date	Evidence	Decision
3	**3-4** Inventory entry and success criteria	Create the agent's inventory entry; write the success criteria, how each will be measured, the current baseline, and the stop criteria that would end the pilot early **Written output:** Inventory entry; success and stop criteria approved by the owner before any run	Page, 'Weeks 3-4: set success criteria before the run': “have the owner approve it before the supervised run starts. If the criteria are set after the results are in, the evaluation measures nothing.” Page, 'Weeks 3-4: set success criteria before the run': “Choose your own thresholds; there is no standard number. Record the current baseline first, since a comparison needs one.” Page, 'Weeks 3-4: set success criteria before the run': “The European Commission's draft Annex 22 on AI, a GMP consultation text from July 2025 that is not in force and does not apply to generative AI or large language models, sets out a useful habit” Page, 'Weeks 3-4: set success criteria before the run': “A discovery pilot does not have to follow it, but the habit carries over.” Page, 'Weeks 7-10: human checkpoints, with the inventory entry made before the run': “Record the agent in your AI inventory before the supervised run, not after.” Page, 'Weeks 7-10: human checkpoints, with the inventory entry made before the run': “Our answer on what an AI agent inventory in pharma R&D should record lists those fields and maps each one to its source, with a free register template.” NIST AI 100-1 (AI RMF 1.0), MEASURE 2.1: “Test sets, metrics, and details about the tools used during TEVV are documented.” NIST AI 100-1 (AI RMF 1.0), MEASURE 2.3: “AI system performance or assurance criteria are measured qualitatively or quantitatively and demonstrated for conditions similar to deployment setting(s). Measures are documented.” NIST AI 100-1 (AI RMF 1.0), GOVERN 1.6: “Mechanisms are in place to inventory AI systems and are resourced according to organizational risk priorities.” European Commission draft Annex 22 - Artificial Intelligence (consultation document, July 2025; not in force), Draft Annex 22, section 4.2: “A process subject matter expert (SME) should be responsible for the definition of the acceptance criteria, which should be documented and approved before the start of acceptance testing.” European Commission draft Annex 22 - Artificial Intelligence (consultation document, July 2025; not in force), Draft Annex 22, section 4.3: “The acceptance criteria of a model, should be at least as high as the performance of the process it replaces.” https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf https://health.ec.europa.eu/document/download/5f38a92d-bb8e-4264-8898-ea076e926db6_en?filename=mp_vol4_chap4_annex22_consultation_guideline_en.pdf https://atlasbiosys.com/downloads/ai-agent-inventory-register-template.xlsx				
4	**5-6** Build and test in a sandbox	Set up an isolated environment with no write access to production systems; log inputs, outputs and model version; test failure cases such as out-of-scope requests, instructions hidden in source documents and missing data; test the off switch **Written output:** Sandbox test log; off switch tested	Page, 'Weeks 3-6: scope the data, then build in a sandbox': “write down who can stop the agent and how its access is withdrawn, and test it once in weeks 5 to 6.” Page, 'Weeks 3-6: scope the data, then build in a sandbox': “OWASP lists LLM06 Excessive Agency” Page, 'Weeks 3-6: scope the data, then build in a sandbox': “source documents can carry text that tries to steer the agent (LLM01 Prompt Injection).” NIST AI 100-1 (AI RMF 1.0), MANAGE 2.4: “Mechanisms are in place and applied, and responsibilities are assigned and understood, to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use.” OWASP LLM06:2025 Excessive Agency (named by the page) OWASP LLM01:2025 Prompt Injection (named by the page) https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf https://genai.owasp.org/llmrisk/llm062025-excessive-agency/ https://genai.owasp.org/llmrisk/llm01-prompt-injection/				

#	Weeks and focus	What gets done / written output (page)	Source (quoted)	Owner	Date	Evidence	Decision
5	**7-10** Supervised run	Run the agent on the real task inside the sandbox; a named scientist reviews every output before it is used; reviewers record accept, correct or reject and why; check the stop criteria weekly **Written output:** Review log; issues list	Page, 'Weeks 7-10: human checkpoints, with the inventory entry made before the run': “That log is the main evidence for the go/no-go decision.” Page, 'Weeks 7-10: human checkpoints, with the inventory entry made before the run': “Write down which outputs need review, who reviews them and how the review is recorded, rather than leaving it to habit.” NIST AI 100-1 (AI RMF 1.0), MAP 3.5: “Processes for human oversight are defined, assessed, and documented in accordance with organizational policies from the GOVERN function.” EMA reflection paper on AI in the medicinal product lifecycle (September 2024), section 2.5.6 (Model deployment): “For all models, especially those where there is no human-in-the-loop, a system risk management plan should be developed that defines likely risks of failure modes of the algorithm. This includes consequences of incorrect predictions/classifications as well as monitoring and mitigation/correction approaches, such as how to trigger a suspension/decommission of the model and how to suspend or decommission it.” https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle_en.pdf				
6	**11-12** Evaluate	Score the results against the criteria fixed in weeks 3 to 4 and the baseline; describe failure patterns; record any change to the criteria, with its reason, rather than rewriting them **Written output:** Evaluation report	Page, 'Weeks 3-4: set success criteria before the run': “If you change a criterion during the pilot, record the change and the reason as a documented deviation from the plan.” The page cites no external source for this step; it follows the criteria fixed in weeks 3 to 4.				
7	**13** Go/no-go gate	Decide to stop, extend the pilot, or continue in the same non-GxP scope; update the inventory entry; route any proposal for GxP use to a separate decision **Written output:** Signed decision record; updated inventory entry	Page, 'Week 13: the go/no-go gate, and a separate gate before any GxP use': “Base the decision on the criteria fixed in weeks 3 to 4, the review log and the issues list, and have the owner sign it.” Page, 'Week 13: the go/no-go gate, and a separate gate before any GxP use': “Stopping is a valid result” NIST AI 100-1 (AI RMF 1.0), MANAGE 1.1: “A determination is made as to whether the AI system achieves its intended purposes and stated objectives and whether its development or deployment should proceed.” https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf https://atlasbiosys.com/downloads/ai-agent-inventory-register-template.xlsx				

#	Weeks and focus	What gets done / written output (page)	Source (quoted)	Owner	Date	Evidence	Decision
8	**Before any GxP use** Separate gate before any GxP use	Before the agent's output is used in a GxP record, process or decision, or in results that may contribute to the evidence presented for regulatory review, treat it as a new intended use: reclassify the inventory entry, record which regime applies (for example GMP, GLP, GCP or GVP) or, where none does, which regulatory expectations apply to that evidence, take it through your quality system's risk assessment and any computerised system validation the applicable regime requires, proportionate to risk, and get it approved under your quality system's procedures before use. **Written output:** None named on the page for this gate; the record is the approval made under your quality system's procedures.	Page, 'Week 13: the go/no-go gate, and a separate gate before any GxP use': “get it approved under your quality system's procedures before use.” Page, 'Week 13: the go/no-go gate, and a separate gate before any GxP use': “A pass at this gate covers only the non-GxP use that was tested.” Page, 'Week 13: the go/no-go gate, and a separate gate before any GxP use': “Pilot results can inform that risk assessment. They count as validation evidence only where your validation procedures allow it, which usually requires that they were generated under conditions representative of the new intended use and documented to the standard those procedures require.” EMA reflection paper on AI in the medicinal product lifecycle (September 2024), section 2.3.1 (Drug discovery): “However, if results contribute to the total body of evidence presented for regulatory review, principles for non-clinical development (see below) should be followed.” European Commission draft Annex 22 - Artificial Intelligence (consultation document, July 2025; not in force), Draft Annex 22, section 1 (scope; GMP): “Following the above, the document does not apply to Generative AI and Large Language Models (LLM), and such models should not be used in critical GMP applications. If used in non-critical GMP applications, which do not have direct impact on patient safety, product quality or data integrity, personnel with adequate qualification and training should always be responsible for ensuring that the outputs from such models are suitable for the intended use, i.e. a human-in-the-loop (HITL) and the principles described in this document may be considered where applicable.” https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle_en.pdf https://health.ec.europa.eu/document/download/5f38a92d-bb8e-4264-8898-ea076e926db6_en?filename=mp_vol4_chap4_annex22_consultation_guideline_en.pdf				