Navigating the EU AI Act

Compliance through governance and observability

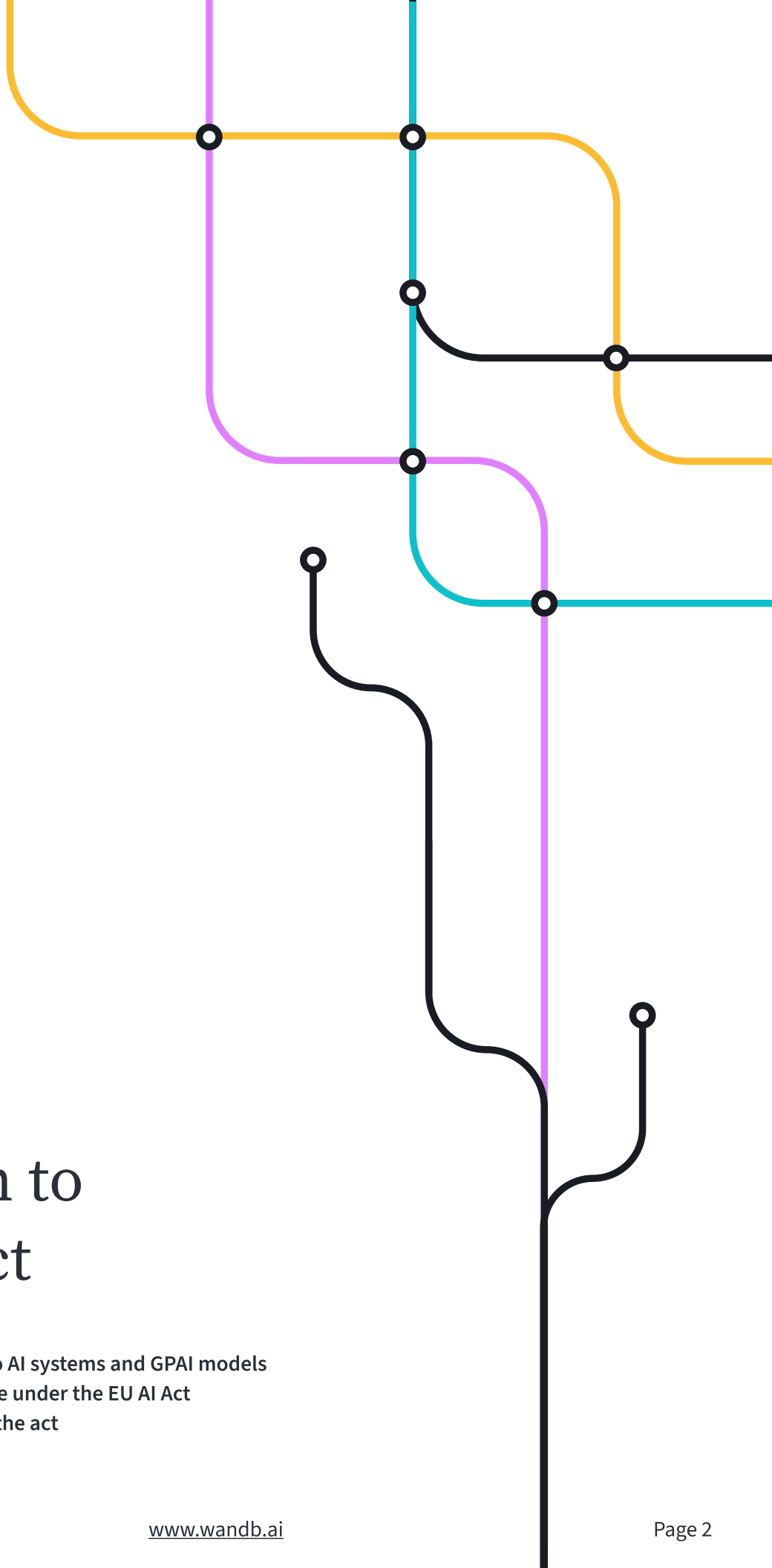
Executive summary

In this white paper we demonstrate how the Weights & Biases platform can help enterprises comply with the requirements for high-risk AI systems under the European Union's Artificial Intelligence Act (EU AI Act).

Enterprises use a wide range of tools to develop, deploy, and govern AI. With the varying obligations defined by the EU AI Act—from stringent requirements for high-risk and general-purpose AI models to voluntary guidelines for low-risk applications—organizations face considerable compliance challenges. There is a need for tools that simplify collaboration, enable standardization of compliant processes, and automate reporting.

Weights & Biases helps enterprises address these complexities by providing powerful observability and governance tools designed to streamline compliance, manage risk effectively, and facilitate the documentation and reporting mandated by the EU AI Act.

In this white paper, we first introduce readers to the EU AI Act (section 1). We then map Weights & Biases core features to the requirements for high-risk AI systems, highlighting how the platform serves both governance and technical teams (section 2). Finally, we illustrate how companies can achieve compliance through an agentic AI case study (section 3). For reference, you can directly explore the **Weights & Biases workspace** and interact with the **generated compliance report** for this case study.



SECTION 1

Introduction to the EU AI Act

Outline:

1.1 EU AI Act's risk-based approach to AI systems and GPAI models

1.2 Obligations enterprises might face under the EU AI Act

1.3 Timelines for the applicability of the act

SECTION 1

Introduction to the EU AI Act

The EU AI Act is a safety legislation designed to promote the adoption of trustworthy AI by ensuring that AI systems do not pose a risk to the health, safety, or fundamental rights of EU residents. It governs both AI systems and General Purpose AI (GPAI) models, and introduces two key features. The first is proportionality: the riskier an AI system or GPAI model is considered under the act, the greater the obligations its operators must meet. The second is distributed obligations: the nature and extent of these obligations vary depending on the actor's role.

1.1 The EU AI Act's risk-based approach to AI systems and GPAI models

a Risk classification for AI systems

The EU AI Act categorizes AI systems into four risk classes depending on their intended purpose: 1) systems with unacceptable risk, 2) high-risk systems, 3) systems with transparency obligations, and 4) low-risk systems.

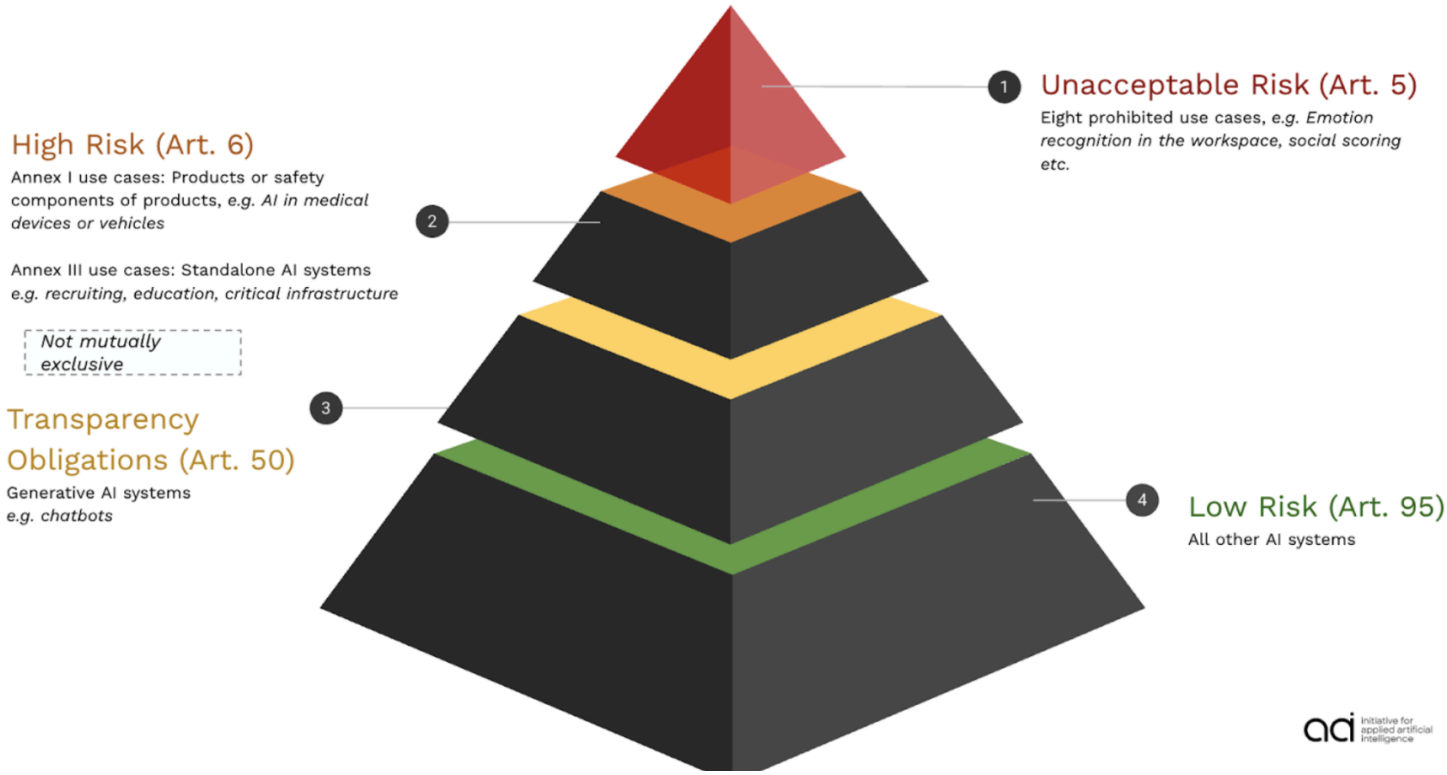


Figure 1. Classification of AI systems depends on the intended purpose

b Risk classification for GPAI Models

Notably—and in deviation from the norm of governing intended use—the EU AI Act mandates special rules for general-purpose AI (GPAI) models. A GPAI model refers to "an AI model that is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks." The obligations for GPAI models are independent of their use and depend on two factors:

Compute

First, the amount of compute used to train these models. Models trained with more than 10^{25} FLOPS of compute are deemed to possess "high-impact capabilities," and thus pose a "systemic risk" to the EU. The EU Commission can also consider additional factors that might indicate whether a model poses systemic risk.

Licence

Second, the licence under which the model is placed on the market. For models below this threshold, the rules depend on whether it was released under an open-source licence or a proprietary licence.

Figure 2. Classification of GPAI models depends on licence and computing power

1.2 Complying with the obligations

The EU AI Act takes a value-chain approach to the obligations for AI systems and GPAI models. Therefore, the obligations depend heavily on an actor's role in relation to the model or system. Although the EU AI Act introduces several roles, including providers, deployers, importers, distributors, and authorized representatives, this white paper focuses on providers. Providers are actors who design and develop an AI system (or have it designed and developed) and place it on the market or put it into service under their trademark. Deployers, on the other hand, are actors who use the AI system under their authority.

	Prohibited	High-Risk	Transparency	Low-Risk	GPAI Model
Provider (builds or sells the system)	Prohibited from making and must withdraw from the market (Art 5)	Product safety requirements for systems and enterprise obligations (Art 8–22)	Transparency obligations (Art. 50(1) & (2))	Voluntary code of conduct (Art 95)	Model governance and transparency depending on systemic risk (Art 53–56)
Deployer (buys or uses the system under their authority)	Must cease operations (Art 5)	Enterprise obligations (Art 26 and 27)	Transparency Obligations (Art 50(3) & (4))	Voluntary code of conduct (Art 95)	--

Table 1. Overview of obligations based on risk class and role

The focus of this white paper will be on the obligations for providers of high-risk AI systems (articles 9, 10, 12, 13, 14, 15) and GPAI models (articles 53 and 55). It is important to note that actors who intend to comply with the requirements for high-risk AI systems and GPAI models must also rely on additional instruments, harmonized technical standards, and the GPAI model Code of Practice (CoP) respectively.

a High-risk systems and the harmonized technical standards

The EU AI Act follows the approach of the EU's new legislative framework for compliance. The act sets down essential requirements for product safety for high-risk AI systems and relies on harmonized technical standards to detail the specific steps that providers must follow to meet the requirements. The standards for the EU AI Act are currently being developed by CEN-CENELEC JTC 21 and are expected to become available later in 2025. You can find more information on the CENELEC website.

b GPAI models and the Code of Practice

Given the relatively novel nature of GPAI models, the EU AI Act creates a unique mechanism for the detailed rules that providers of such models must follow: The GPAI model Code of Practice (CoP). The CoP will play a similar role for GPAI model providers as technical standards do for high-risk AI systems. These codes are currently being drafted and will become available in May 2025. You can find more information about the CoP on the official website.

1.3 Timelines

While the EU AI Act entered into force on August 2, 2024, the dates from which operators of AI systems or GPAI models must comply with the rules will apply in a staggered manner based on the risk class of the system. The relevant implementation periods are as follows:

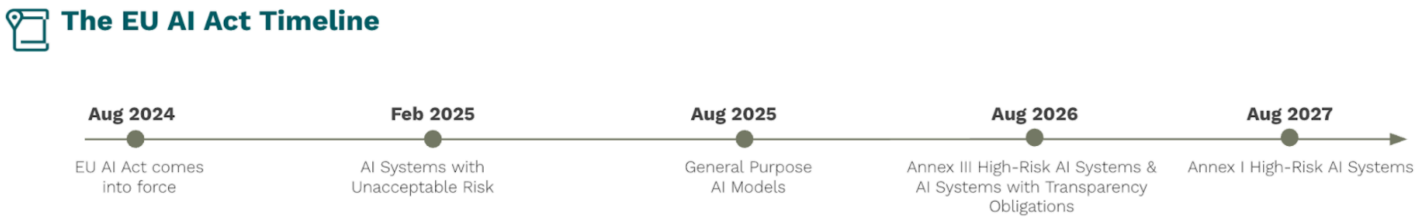
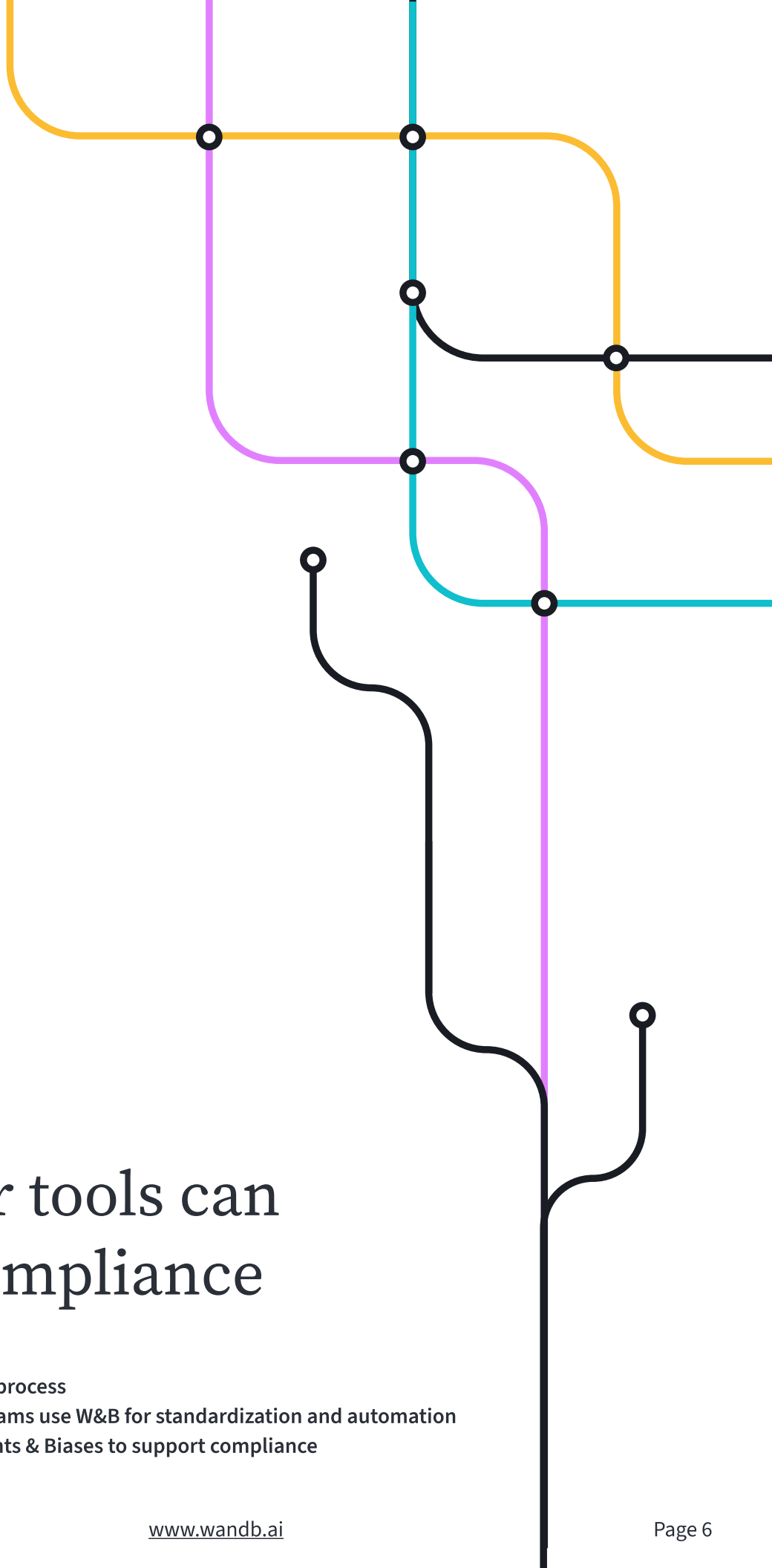


Figure 3. Implementation timeline for the EU AI Act



SECTION 2

AI developer tools can help with compliance

Outline:

2.1 Typical enterprise AI compliance process

2.2 How platform and engineering teams use W&B for standardization and automation

2.3 How governance teams use Weights & Biases to support compliance

SECTION 2

AI developer tools can help with compliance

The EU AI Act introduces complex requirements, particularly for high-risk AI systems, that many organizations struggle to meet. Different stakeholders operate on disconnected platforms, with no common translation between regulatory and technical needs. Teams follow diverse processes and goals, while reporting and enforcement remain largely manual, error-prone, and time-consuming. These challenges highlight the urgent need for integrated enterprise processes and tools to ensure efficient and consistent compliance.

a Different stakeholders

Governance, platform, and engineering teams have their own systems. There's no translation between regulation, platform, and engineering requirements and results.

b Different processes

Various teams have their own processes with different goals. Standards are only considered per team and results and artifacts are not centrally shared.

c Manual reporting

Regulation checklists are manually created and filled out. In a highly empirical and iterative development process, this leads to errors and a lot of time overhead.

d Manual enforcement

Standards and requirements checked manually in retrospect. Multiple parallel model training, deployment, and stakeholders makes this challenging.

2.1 Typical AI compliance process adopted by enterprises

Organizations typically structure compliance responsibilities across three key teams: the AI governance team, the platform team, and engineering teams. EU AI Act compliance is an iterative process with feedback loops for refinement, requiring effective collaboration between these teams. Figure 4 illustrates the functions of these teams and their interactions for compliance. For more details on these processes, see the upcoming appliedAI Initiative white paper "Designing high-risk AI systems under the AI Act."

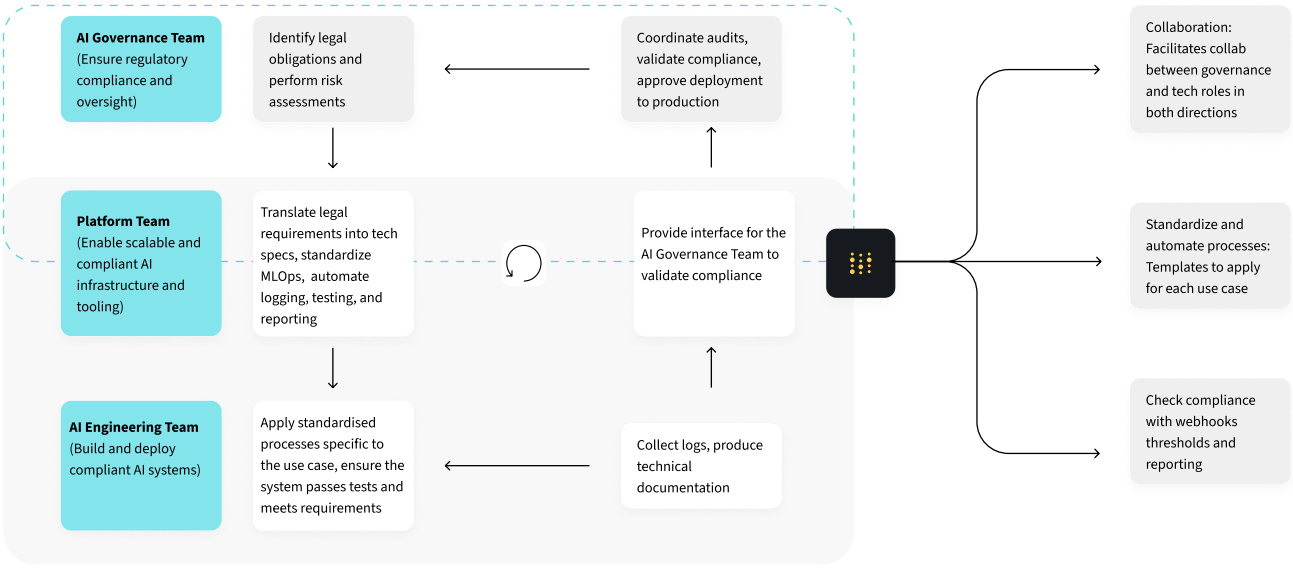


Figure 4. Interactions between enterprise teams for EU AI Act compliance

2.2 Weights & Biases supports platform and engineering teams

In this section, we highlight how Weights & Biases can help platform and engineering teams establish processes and workflows to comply with requirements for high-risk AI systems (articles 9, 10, 12, 13, 14, 15) and GPAI models (articles 53 and 55). We summarize key requirements of the EU AI Act and map them to Weights & Biases features.

ARTICLE 9

Risk management system (identify, evaluate, and mitigate risks)

Requirement	Solution
Eliminate or reduce risks through “adequate design and development”	• W&B Weave provides a flexible evaluation framework and scorers to measure and reduce quality and safety risks of AI applications. • Human-in-the-loop (HITL) integration through user and expert feedback further mitigates risks and aligns AI with user needs. • W&B Artifacts tracks data lineage and metrics, mitigating bias and imbalances.
Implement “adequate mitigation and control measures”	• W&B Weave Guardrails protect against harmful content reaching users. • W&B Registry tracks models, datasets, and metadata for reproducibility and troubleshooting to reduce the risk of inaccurate outputs. • W&B Models tracks experiments while training and fine-tuning models.
“Provide information” and “adequate training” for deployers	• W&B Reports helps developers document experiment metrics and insights. • We offer free courses through the Weights & Biases AI Academy and the AI Master Class to help developers get educated on AI systems.

Feature highlights

W&B Weave

- Generate custom metrics to evaluate and monitor AI systems.
- Versioned datasets with critical test cases mitigate regression risks.

W&B Models Experiments

- Evaluate model performance under varying scenarios.
- Log different model versions and associated test results.

W&B Registry/Automations

- Execute predefined tests on new model and data versions, supporting continuous monitoring requirements of Risk Management System (RMS).

Safety Scorers	Toxicity | Bias | PII detection | Hallucinations
Quality Scorers	Coherence | Fluency | Context relevance

ARTICLE 10

Data and data governance

Requirement

Training, validation, and test data must satisfy governance and quality requirements before deployment

Solution

- Weights & Biases enables customers to create and maintain high-quality datasets for training, fine-tuning, and testing both models and systems.
- W&B Weave helps customers identify and resolve dataset issues, collect real-world examples from production, and generate expert human labels.

Feature highlights

W&B Weave

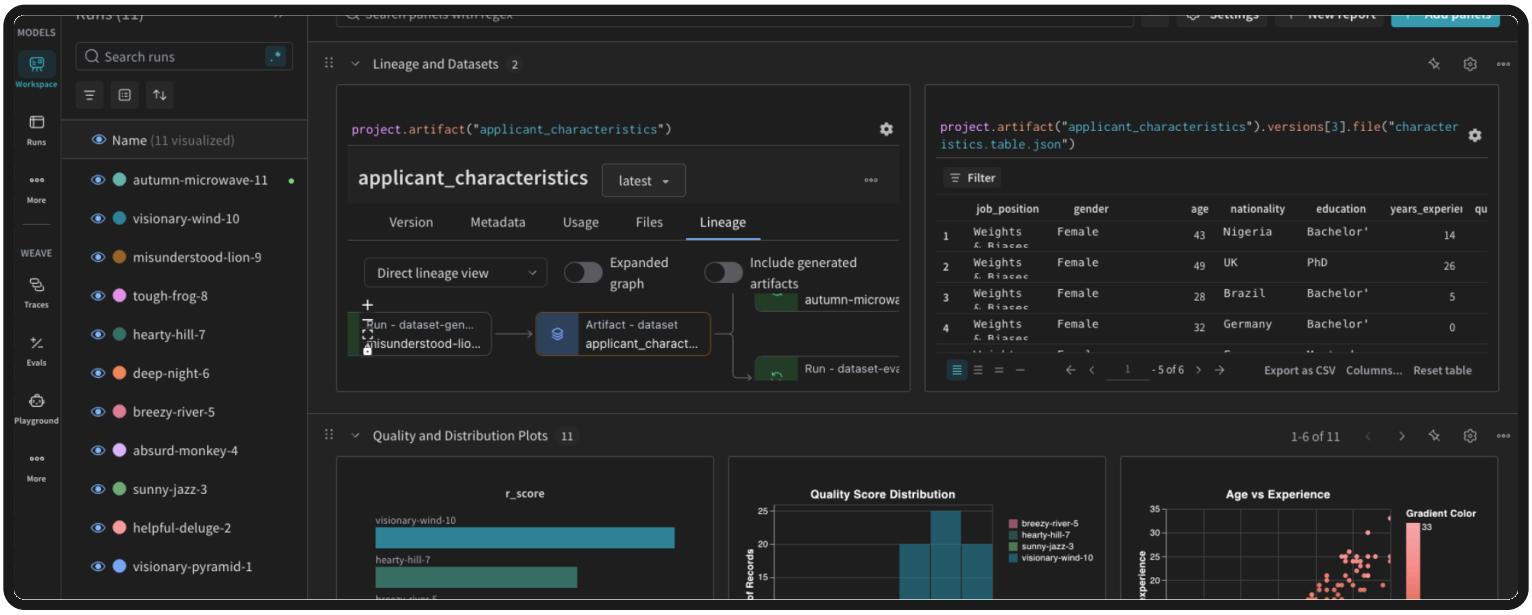
- Create high-quality datasets from production traces.
- Edit and manage datasets to enhance quality.
- Collect expert annotations through structured user interfaces, improving quality of labeling.
- Upload offline data to build comprehensive datasets.

W&B Models Experiments

- Visualize data to validate assumptions and identify issues.
- Log end-to-end data transformation and testing.
- Identify and mitigate bias using external libraries or custom methods.
- Prepare custom metrics and tests to meet data quality requirements.

W&B Registry/Automations

- Build predefined data quality tests.
- Standardize and enforce testing and evaluation procedures, including role-based access control (RBAC) for execution.
- Support continuous integration and continuous deployment (CI/CD) for ML
- Automatically trigger training, evaluation, and deployment jobs.



Lineage graph, datasets, and data distribution plots

ARTICLE 12

Record keeping

Requirement

"Technically allow for the automatic recording of events (logs)" to ensure traceability of risks and compliance in production

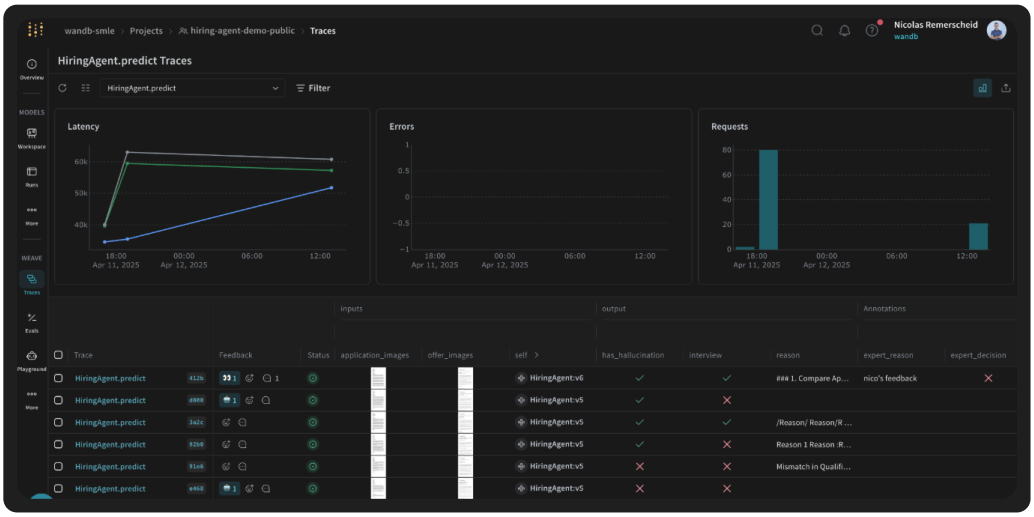
Solution

- W&B Weave Traces automatically log all inputs, outputs, code, and metadata in an AI application at a granular level.
- Custom scorers look for specific events and record them in traces.
- Traces can help you monitor, identify, and mitigate risks during development and post-production for compliance.

Feature highlights

W&B Weave

- Log traces from any AI application in production or testing.
- Live score production traces to detect key events.
- Track metadata, datasets, and model version logs.
- Automatic versioning to keep track of changes.
- Comparison capabilities to see impact of changes.



Production traces and aggregated metrics over time

ARTICLE 13

Transparency and instructions of use

Requirement

Provide clear usage instructions so that the AI system is sufficiently transparent for deployers to understand and use it appropriately

Solution

- Use W&B Weave to run evaluations and W&B Models to run experiments to baseline the performance of AI systems and models.
- W&B Registry allows developers to publish and share models and their performance baselines and any metadata such as hyperparameters to create a system of record for transparency.
- W&B Reports help document metrics and characteristics about model performance to provide instructions of use to downstream actors.
- 3rd party libraries can be integrated for interpretability of AI systems.

Feature highlights

W&B Weave

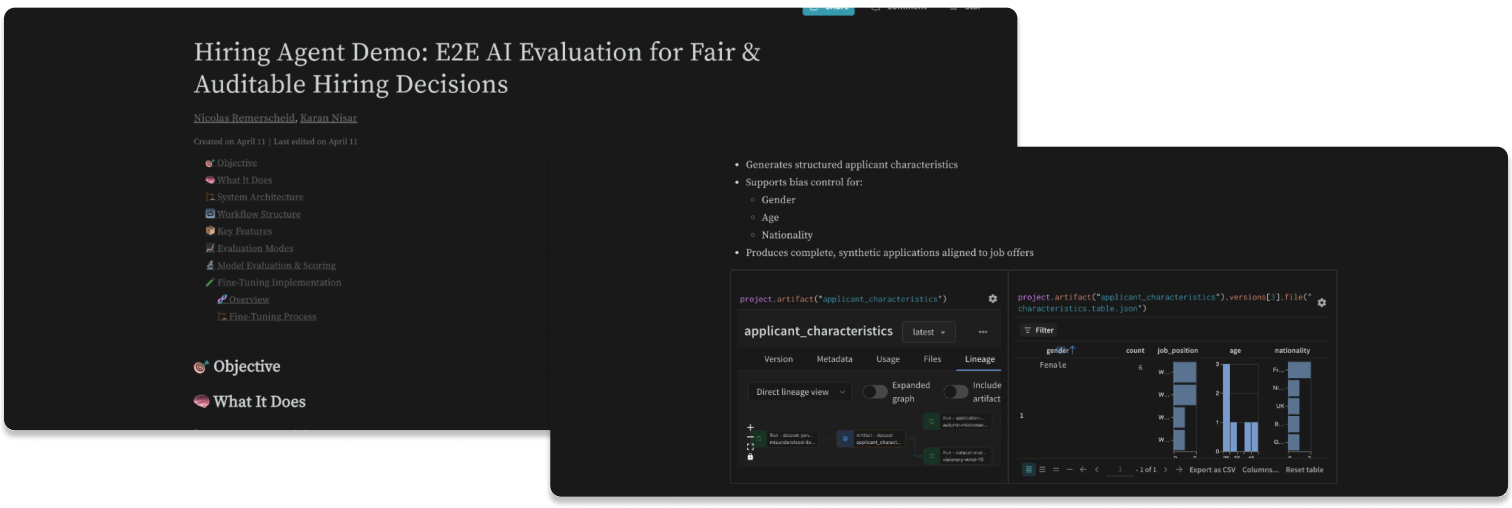
- Centrally track all evaluation data to enable reproducibility, collaboration, and governance.
- Trace lineage back to LLMs used in your application to make continuous improvements.
- Weave automatically versions code, datasets, and scorers by tracking changes between iterations, enabling you to pinpoint performance drivers.

W&B Models Experiments

- Enable versatile quantitative and qualitative analysis through W&B Plots and W&B Tables.
- Easily create custom charts or visualize explainability/interpretability plots directly through integrations with popular frameworks.
- Use W&B Sweeps to run hyperparameter optimizations with automated visualizations of feature importance and correlation.

W&B Registry/Reports

- Create and track custom model and data cards.
- Evaluation results and benchmark baselines can be shared with downstream users for auditing and governance.
- A system of record offering easy access to AI artifacts such as models and datasets and detailed lineage tracking allows you to rebuild any model and reproduce any task in the ML lifecycle.



Report with plots that dynamically update when data changes

ARTICLE 14

Human oversight

Requirement

AI systems must be designed such that a human can observe their functioning and control them when necessary

Solution

- W&B Weave Monitors allow developers to detect failure modes, anomalies, and system malfunction. They can write custom logic to hand off control to a human overseer.
- Human operators can be trained on the system baseline performance, limitations, and characteristics using Registry, Reports, and Evaluations in W&B Models and W&B Weave to take over control when necessary.

Feature highlights

W&B Weave

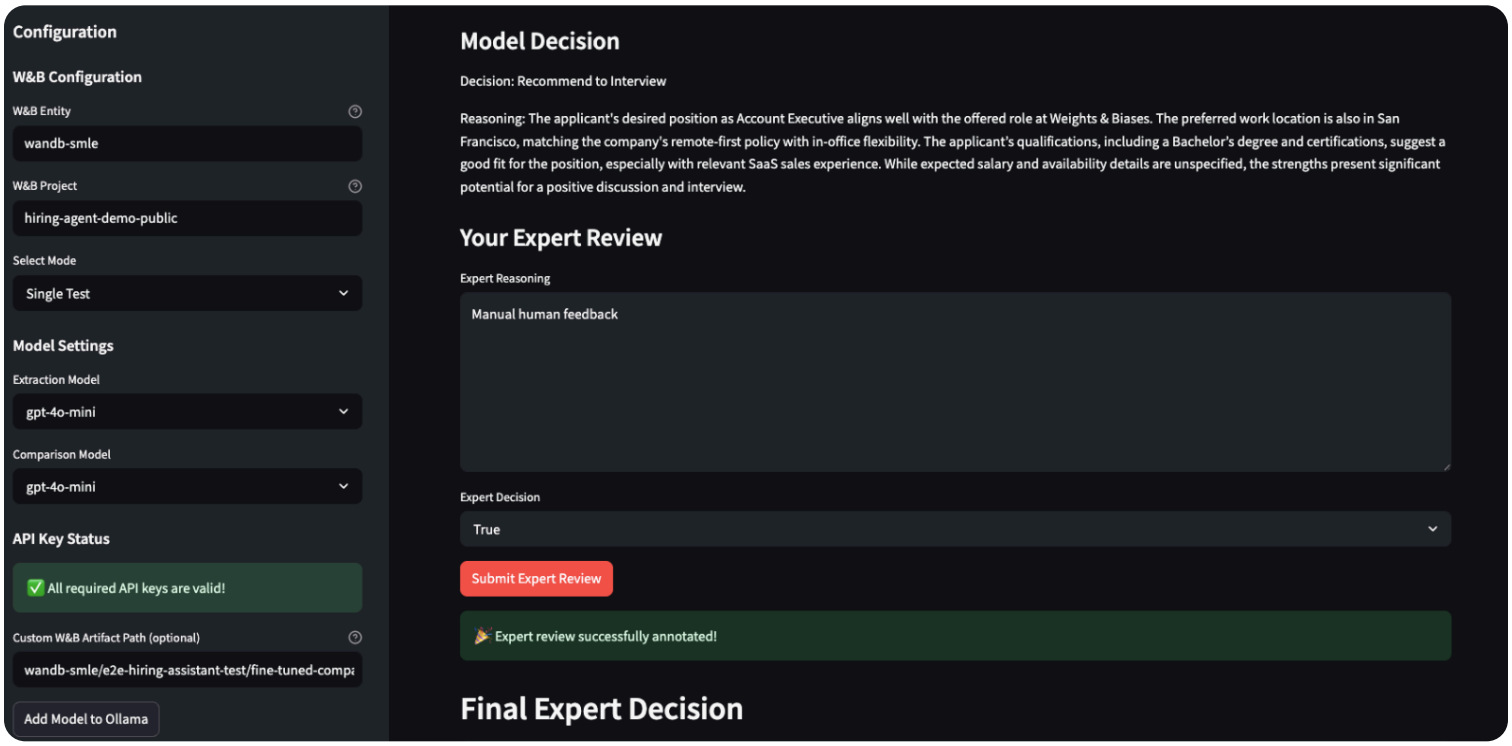
- Monitors can be used for alerting. Guardrails detect issues and choose alternate handling such as routing to humans for intervention.
- Human annotations allow expert grading of responses.

W&B Models Experiments

- Trigger alerts based on configured thresholds.
- Users can access all the logged data in Weights & Biases via the API.

W&B Reports

- Downstream actors can view a public report with embedded data and model benchmarks.
- Human operators can be trained using technical reports from training runs.



Custom interface for human experts to override agent decisions

ARTICLE 15

Accuracy, robustness, and cybersecurity

Requirement

Ensure AI systems meet accuracy, robustness, and cybersecurity requirements before deployment and maintain them during operation

Solution

- W&B Weave, W&B Models, and W&B Registry provide a rigorous tool set to measure, track, and document accuracy, robustness, and safety.
- W&B Weave provides post-production guardrail fail-safes to detect and handle errors, inconsistencies, security issues such as PII leaks, and cyber attacks such as prompt injection.

Feature highlights

W&B Weave

- Run evaluations to baseline key metrics and monitor their trend over time in production to maintain consistency and robustness.
- Implement guardrails to detect and mitigate malicious activities and harmful content.

W&B Models Experiments

- Log model performance metrics to measure multiple aspects of performance and accuracy.
- Import libraries to test for robustness.
- Integrate with more specialized testing tools (e.g. Giskard, for which Weights & Biases provides a report).

W&B Automations

- Automate and standardize tests on new models and data.
- Trigger evaluation pipelines automatically in critical lifecycle stages to prevent model regressions during development.

Report comparing evaluations based on a curated test dataset

ARTICLE 53

Obligations for providers of GPAI models

Requirement

Respect text/data mining rules, publish summary of training content, and, if providing proprietary model, submit training and eval results to the AI Office and inform downstream actors about the model

Solution

- W&B Models enables model providers to log datasets along with models in the Registry to meet regulatory reporting obligations relating to training data.
- W&B Reports can be used to document and publish this information dynamically for downstream actors to understand capabilities and limitations of the model and to comply with obligations.

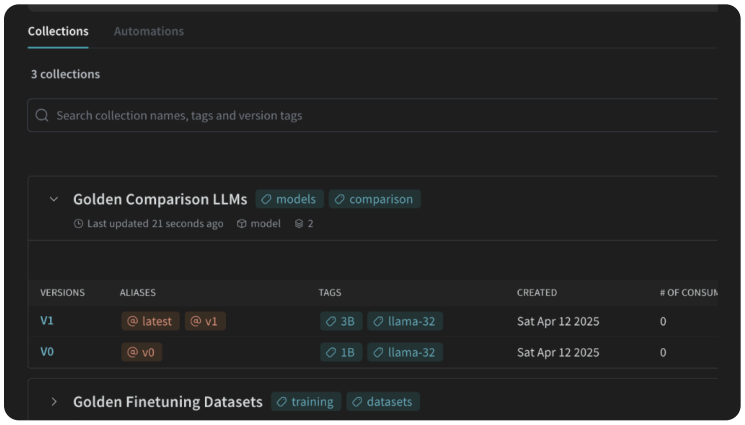
Feature highlights

W&B Models

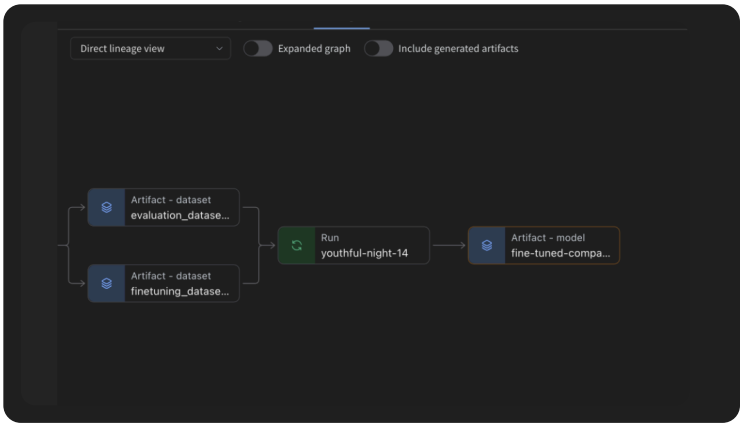
- Log model evaluations and test results centrally for reproducibility.
- Log model and dataset versions automatically for auditing and compliance purposes.
- Reports help document findings from machine learning experiments and share them with stakeholders dynamically.

W&B Weave

- Describe and log inputs to and outputs from the AI system for complete end-to-end traceability.
- Run evaluations and log results in a central place.
- Automatic versioning of models, prompts, metadata, and code for change tracking and governance.



Org-level model and dataset registry with granular RBAC and automations



End-to-end data and model lineage

ARTICLE 55

Obligations for GPAI models with systemic risk

Requirement

Perform evaluation in accordance with standardised state-of-the-art protocols and tools, assess and mitigate possible systemic risks at Union level, document and report serious incidents, and ensure adequate cybersecurity protection

Solution

- W&B Models allows model providers to evaluate models systematically, track the results centrally, and report key findings to regulatory bodies.
- W&B Weave allows customers log every LLM call and flag serious incidents for documentation and reporting purposes.

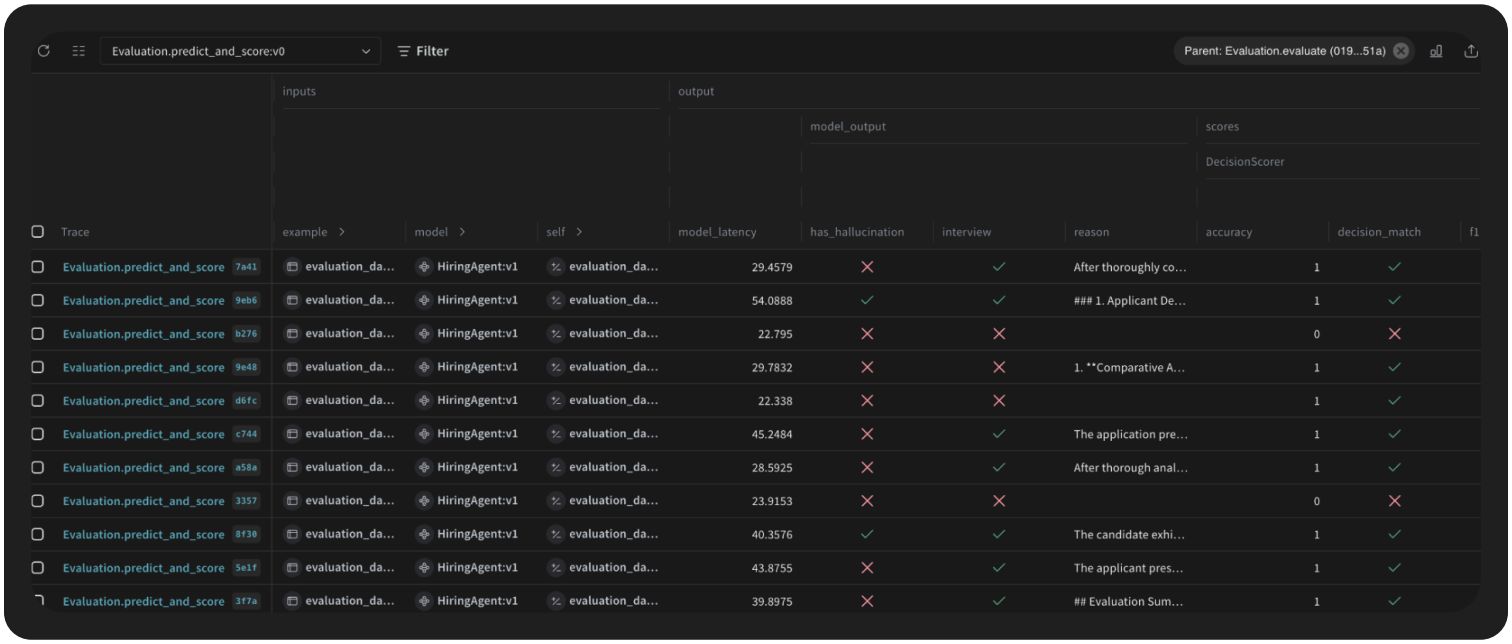
Feature highlights

W&B Models

- Perform and log model evaluations.
- Implement and assess mitigation measures.

W&B Weave

- Log outputs in order to assess incidents.



Scalable agent tracing, consumable through UI or API

2.3 Weights & Biases empowers governance teams

Most enterprises have governance teams that enforce compliance and verify reporting. This is critical because compliance with the EU AI Act requires formal reporting for both high-risk AI systems and GPAI models.

a Technical documentation under the EU AI Act

Under Article 11 and Annex IV of the EU AI Act, providers of high-risk systems must prepare detailed documentation, both about the architecture of the AI system and evidence proving that they complied with the high-risk requirements. This documentation, along with the quality management system, is the basis for assessing the conformity of an AI system.

Providers of GPAI models, meanwhile, must prepare documentation for the AI Office and for downstream actors under Annex XI and XII. The documentation of GPAI model providers will be assessed by the AI Office to check for conformity.

In the table below, we summarize some of the key reporting requirements that are related to Articles 9-15 and GPAI models (note that this list is not exhaustive, see Article 11 and Annex IV for more information).

Annex IV	Custom report for a general and detailed description of the AI system
Annex IV & Article 9	Detailed description of the risk management plan
Annex IV & Article 10	Data sheets describing data acquisition, processing, provisioning, and other governance and quality activities
Annex IV & Article 12	Information about monitoring the AI system
Annex IV & Article 13	Assessment of instructions of use
Annex IV & Article 14	Description of human oversight functions
Annex IV & Article 15	Description of appropriateness of performance metrics for the specific AI system along with cybersecurity measures put in place
Annex XI and XII and CoP GPAI Models	General description of the model Description of how the model can be integrated into a system along with additional information, such as training methods, evaluations, and resource consumption For models with systemic risk, detailed description of risk evaluation and mitigation strategies

b How Weights & Biases empowers governance teams

Weights & Biases offers a system of record for AI models and applications that all of the stakeholders mentioned above can use as a central platform to track and share data, which can then be used to generate the documentation required by the EU AI Act. Since the EU AI Act does not specify formal rules for preparing this documentation, it is essential for companies to develop efficient implementation strategies. With W&B Reports, predefined templates can be established to standardize the compliance process, while W&B Automations enables predefined compliance workflows that run automated tests and generate technical documentation. Other stakeholders can easily access the resulting technical documentation or AI system details for audits and external reporting.

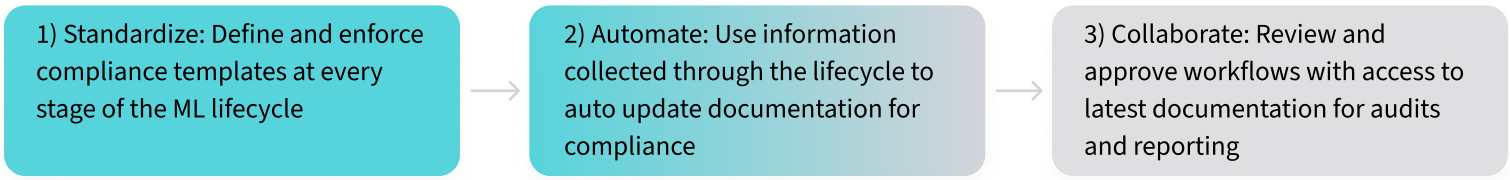
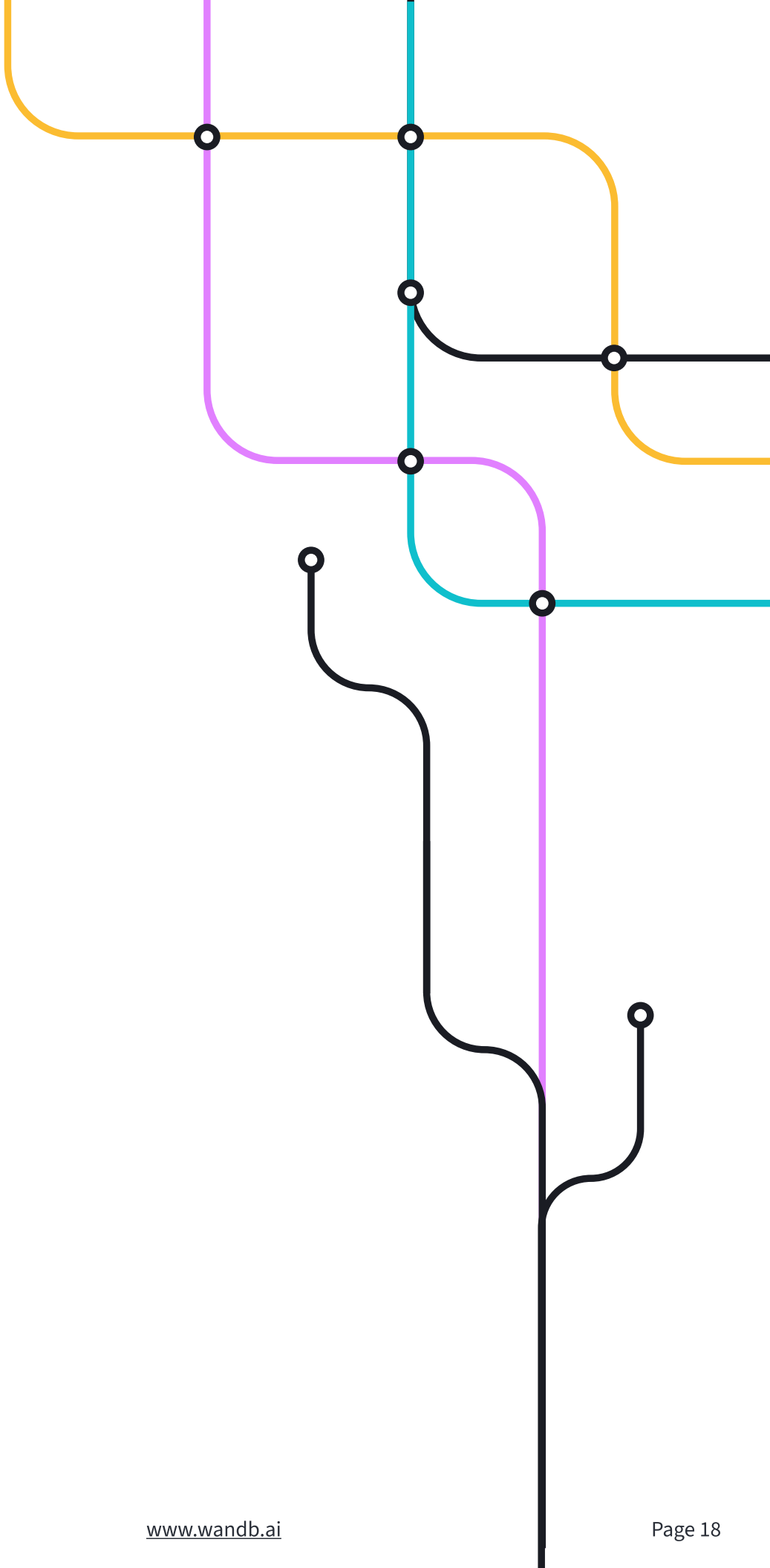


Figure 5. Best practice process for compliance

Steps	Jobs to be done	Solution
1 Standardize	Establish compliance checks at different stages of the ML lifecycle, pre-defining the information to be generated.	W&B Models integrates with orchestration tools (Argo, Dagster, and Prefect) to enforce compliance workflows throughout the ML lifecycle.
	Until the EU Commission or national authorities standardize templates for documentation, companies should use existing tools like Google's Data Cards and Meta's System Cards.	W&B Reports can follow pre-defined documentation templates to ensure alignment with emerging regulatory standards and consistency across projects.
2 Automate	Ensure adherence to development and reporting best practices (first line of defense) to implement robust development-to-production pipelines.	W&B Automations can run automated tests before production deployment, validating compliance with predefined thresholds throughout the ML lifecycle.
	Auto-update compliance documentation using information collected in each step of the lifecycle.	W&B Automations enable real-time updates to reports, ensuring governance teams always have the latest compliance data.
3 Collaborate	Enable management oversight for compliance validation (second line of defense).	W&B Reports enable governance teams to dynamically visualize and access compliance documentation.
	Facilitate internal and external audits (third line of defense).	With W&B Reports, key stakeholders can access the generated technical documentation and AI system details for audits and external reporting.



SECTION 3

Case study

SECTION 3

Case study

3.1 Planning

The following case study provides a real-world example illustrating how Weights & Biases helps organizations meet the technical requirements of the EU AI Act while facilitating collaboration between technical and non-technical stakeholders.

We introduce a fictional company, Acme Inc., which is designing an AI solution for its internal HR team. The intended purpose of the AI system is to evaluate whether a job candidate should be interviewed for an open position based on a comparison of their resume and a description of the position.

3.1.1 Use case description

Based on the intended purpose, the engineering and platform teams proposed to design an AI System composed of two models in cascade as seen in Figure 6:

- An extraction model (e.g. GPT-4o mini): It parses CVs and positions into tabular data.
- A comparison model (e.g. GPT-4o): Compares CVs and positions.

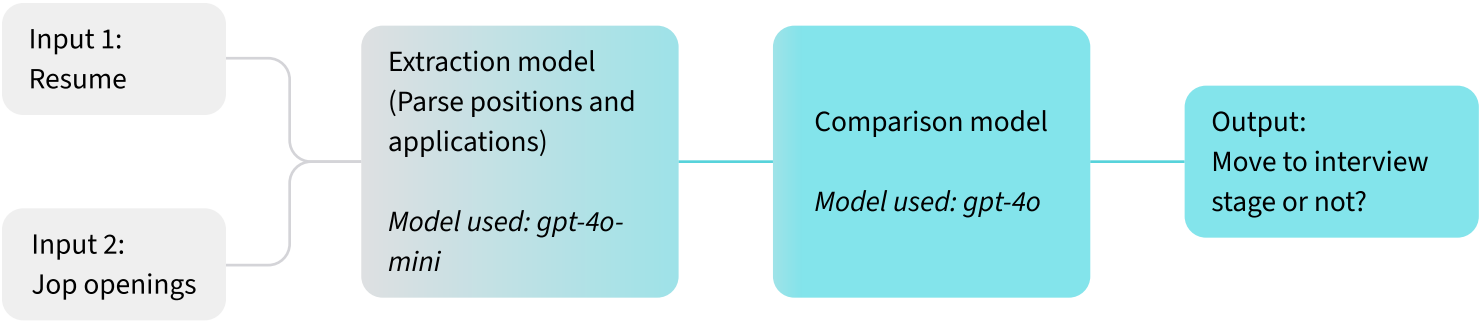


Figure 6. AI system boundaries for the case study

3.1.2 Identify legal obligations

Once a system's intended purpose and architecture have been defined, it is important for an enterprise to identify the obligations they must meet. There are two considerations here:
1) What are the obligations for the AI system?
2) Are there any obligations to be met at the level of the GPAI models (in our use case, GPT-4o and GPT-4o-mini)?

Note: It is important to remember that GPAI models and AI systems are two different concepts under the EU AI Act. Typically, providers of GPAI models like OpenAI, Anthropic, and Meta must meet the obligations for GPAI models while downstream actors who integrate those models must meet the obligations for AI systems. However, the Commission will also create rules for downstream actors who fine-tune or modify models.

In our case study, the following obligations apply:

1. **AI system level obligations:** After applying the risk classification procedure to the AI system, we identify it as a high-risk AI system. Specifically, Article 6(2) and Annex III(4)(a) deem systems that are intended to be used for the recruitment or selection of natural persons as high-risk. This implies that the provider of such a system, in our case Acme, must meet the obligations set out in Chapter 3, Section 2 (see Table 2) while building the AI system.
   Disclaimer: There are additional obligations for providers of high risk systems in Chapter 3, Section 3 that are not related to technical workflows such as Quality Management System (QMS), registering high-risk AI systems, and cooperating with competent authorities. These obligations are out of scope of this white paper.

2. **Model level obligations due to fine-tuning:** In our case study, we start with two models: GPT 4o and 4o mini. Open AI, as the provider of these GPAI models, would have to meet the obligations listed in section 2.2. On the other hand, Acme also considers fine-tuning open-source models. Therefore, Acme's obligations are limited to fine-tuning or modification of the base model. Given that the EU Commission will likely provide more details about this at a later date, the obligations related to fine-tuning are excluded from the scope of this white paper.

3.1.3 Scoping technical requirements

Once the obligations have been identified, it is important to translate the legal requirements into technical requirements. For our case study, we group the requirements for high-risk systems into three clusters of activities:

- First, the standardized activities that are required for providers of high-risk AI systems under Article 10 (data governance), Article 12 (event logging), Article 13 (transparency and instructions of use), Article 14 (human oversight) and Article 15 (accuracy, robustness, and cybersecurity).
- Second, additional risk mitigation techniques under Article 9 that providers of AI systems have to identify themselves based on the risk management system (identifying, evaluating, mitigating, and testing). Risk management is typically a lengthy process. For the purpose of brevity, we exclude a description of the full process for identifying and estimating risks. Instead, we briefly describe risks and mitigation techniques that are most relevant to our case study in the following table.

Risks	Mitigation techniques	Proposed testing
Biased comparisons due to unrepresentative data for the general population	Generate synthetic data with appropriate data quality checks	Quantitative scores on test dataset and qualitative manual assessment
Biased comparisons due to unrepresentative test dataset for production	Update test dataset continuously based on annotated production traces	Measure of data drift over time
Arbitrary decision making due to hallucinations from underlying models	Use a hallucination guardrail and fine-tune the comparison model	Guardrail efficacy versus human expert and score on test dataset
PII leakage either to model provider or unauthorised individuals	Use a guardrail to mask PII during production usage	Guardrail efficacy on PII masking dataset (not detailed in this paper)

Applying the identified mitigation techniques for our AI Hiring Agent, we amend the original use case description (in section 3.1.1) in the following ways. We introduce:

a. **A fine-tuned comparison model (Llama 3.2, fine-tuned):** Generates a structured output with a binary decision and textual reason based on the input prompt with the extracted information. We fine-tune a much smaller open-source alternative to compare it to the GPT-4o model as a baseline. We deem fine-tuning a smaller open-source model as sufficiently promising to improve the alignment of the decision reasoning with our company policies and decrease usage costs in production.

b. **Additional guardrail model (gpt-4o-mini):** Compares the generated hiring reason with the application and job position. If the reason doesn't directly follow from either the application or the job description, it flags a hallucination and gives feedback to the comparison model, which tries again (self-reflection). If the model still hallucinates, the agent reaches out to a human expert operator who can manually input the decision and reason through an operator UI.

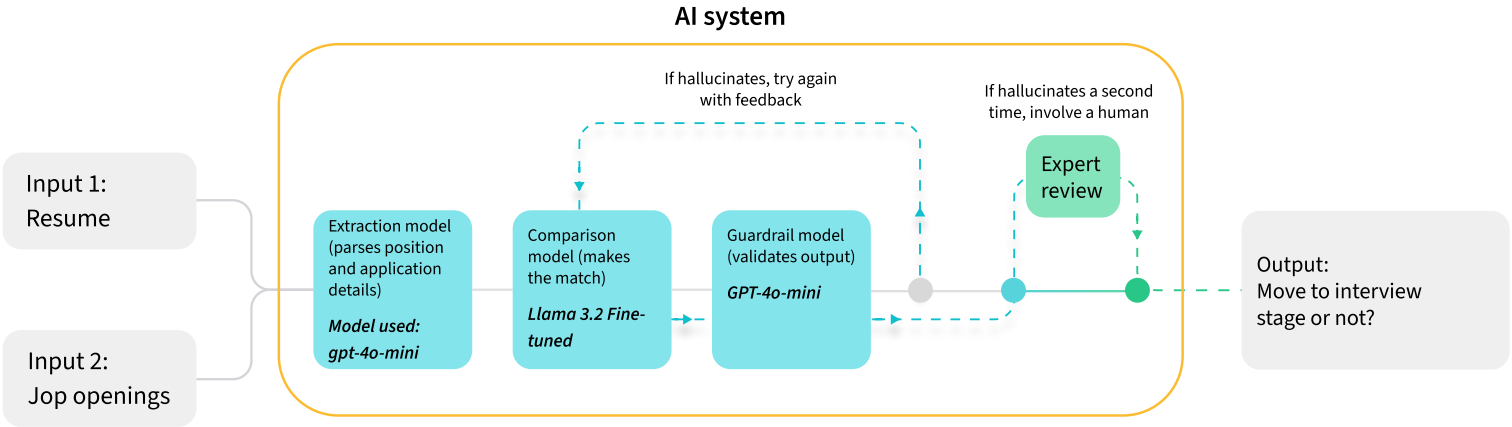


Figure 7. AI system boundaries for the case study after adding a guardrail model

- Third, in addition to standardized activities and risk mitigation techniques, enterprises also need to compile technical documentation describing the architecture of the system and evidence of compliance. For our case study, we identified the activities Acme must perform based on the legal requirements in Table 7.

In the table below, you can find an overview of all the obligations mapped to the ML lifecycle (MLLC). The MLLC is a cyclical process used to develop, train, and deploy AI models, utilizing the data available from the corresponding applications (for more details, see the MLLC definition).

Table 7. Mapping technical requirements to the MLLC

	Data engineering	Modeling	Deployment and monitoring
Standardized activities (Article 10-15)	• Data collection process via W&B BYOB and reference artifacts • Data exploration and preparation via W&B Tables and W&B Models • Data quality evaluation using W&B Reports together with W&B Experiments, and W&B Artifacts with versioning and lineage • Automating repetitive workflows from the W&B Registry with W&B Automations and W&B Webhooks	• Model fine-tuning and hyperparameter optimization using W&B Models Experiments and W&B Models Sweeps • Performance evaluation via W&B Weave Evaluations and W&B Weave Tracing • Versioning of models, datasets, and prompts using W&B Weave	• Development to production hand-off via W&B Reports • Traceability and logging via W&B Traces • Privacy and data protection via W&B Weave's PII masking function • Human oversight interface via W&B Weave API • Continuous performance evaluation via W&B Weave Monitors and W&B Weave Annotation capabilities
Additional risk mitigation techniques (Article 9)	• Create representative synthetic dataset through multi-step generation process to address data bias	• Robustness and risk mitigation via Weave Guardrails • Fine-tuning the comparison model on sample resume-job pairs to enhance decision-making and reasoning performance	• Annotated production traces • Implement W&B Weave's PII masking setting to prevent sensitive data being logged to Weave and automated post-processing decorator to filter data being sent to LLM providers
Technical documentation (Article 11)	Use W&B Reports to generate technical documentation and instructions of use (manually via the UI or programmatically using script templates)		

3.2 Execution

The execution phase implements the technical requirements defined in the planning phase across the various stages of the ML lifecycle. Below, we summarize how Weights & Biases can support the execution phase. For detailed technical implementation and further explanation, refer to the **compliance report** generated by Weights & Biases for this project and the associated W&B Workspace.

3.2.1 Data engineering

Weights & Biases helps organizations align early ML lifecycle practices with the data and governance requirements outlined in Article 10 of the EU AI Act. Following are several capabilities offered by Weights & Biases to support compliance with the EU AI Act.

Data collection

To evaluate and fine-tune the model, we used Weights & Biases bring-your-own-bucket (BYOB) connection to access proprietary data. The BYOB allows us to store the data in a connected object storage (this can be Amazon S3, GCS, Azure Blob, etc.) and pull any information and metadata about the data for documentation.

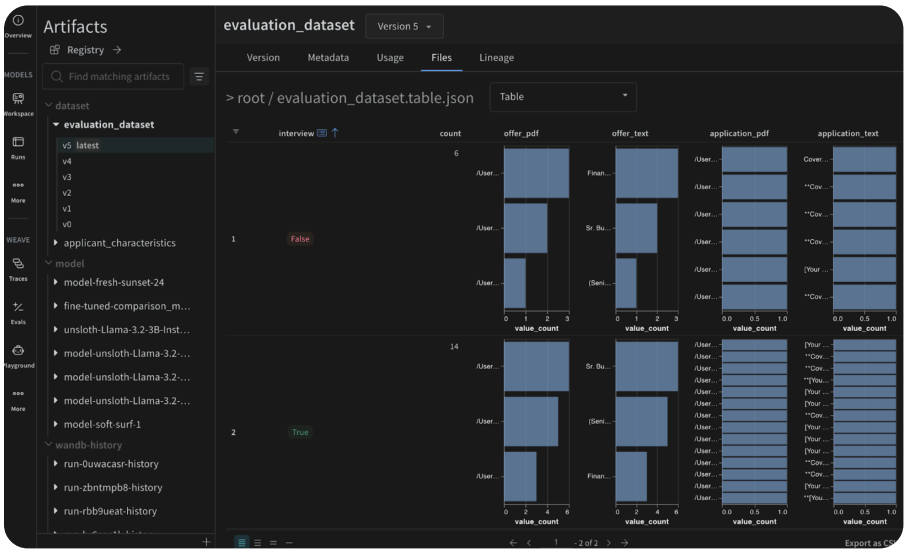


Figure A. EDA reveals unbalanced examples in the [dataset]

Data exploration and preparation

To check for data quality and biases, we conducted an exploration of the dataset distribution using Tables in W&B Models (see Figure A) and found that the data is under-representative. To address this risk, we generate synthetic data of applicants.

Data lineage

The entire data preparation process described above can be easily traced back using the lineage graph automatically generated by W&B Models (see Figure B), improving reproducibility and facilitating root-cause identification and documentation.

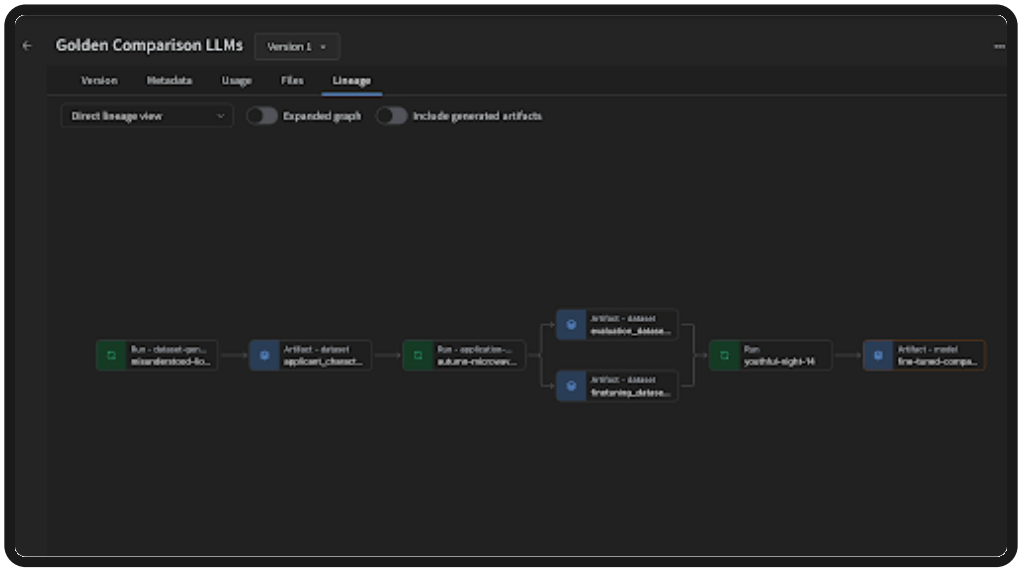


Figure B. Lineage of the [two-step dataset generation process] (green experiments, blue artifacts)

Data quality report

Based on this synthetic data, we can now calculate a data quality score by evaluating the five data quality characteristics specified in the EU AI Act: representativeness, statistical properties, completeness, relevance, and error-free (see Figure C). We then gather all plots and information in a central data quality report using W&B Reports (see also the generated compliance report).

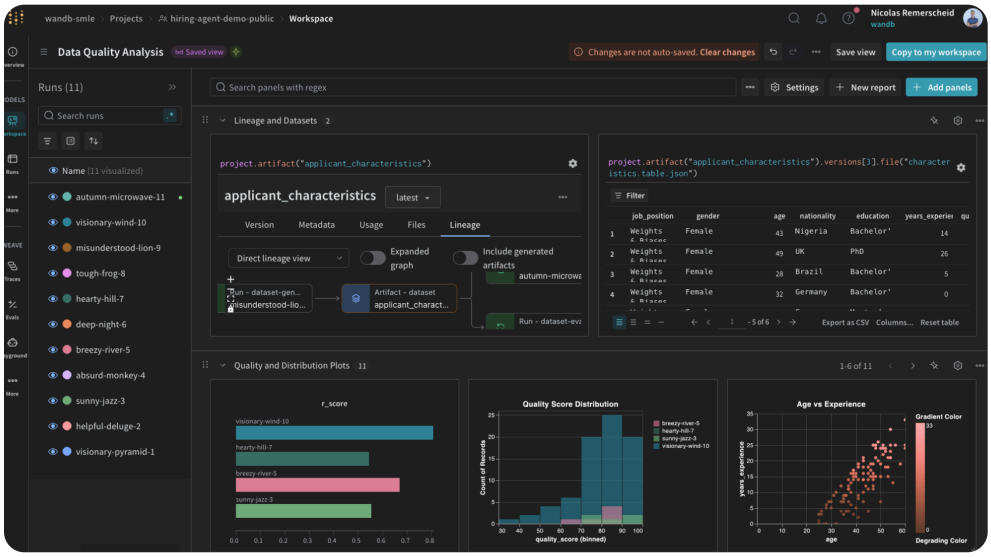


Figure C. Data quality calculation based on the R-Score

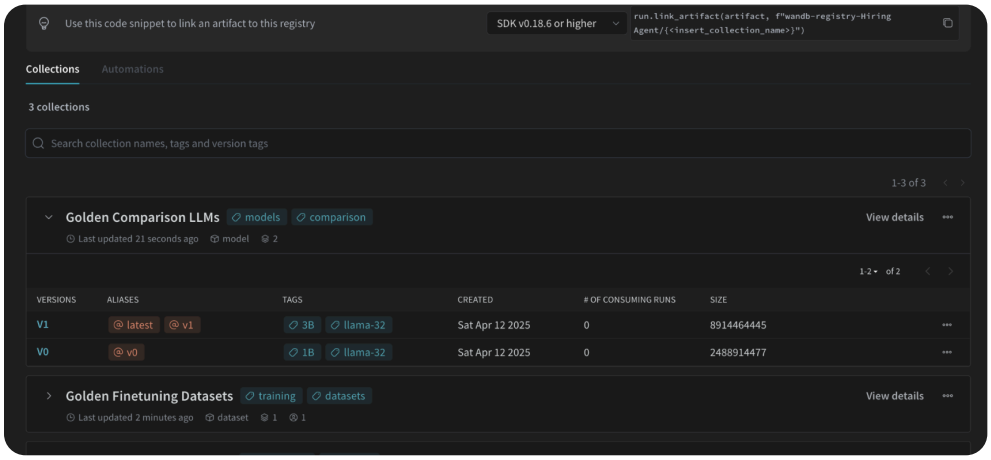


Figure D. W&B Registry is the central place to store the best datasets to be consumed by downstream users

W&B Registry and Automations

Once the quality of the generated dataset is validated, it can be shared with downstream teams through the W&B Registry (see Figure D). This process of generating the dataset, validating its performance, and pushing it as the new gold standard on the Registry can be also automated using W&B Automations.

3.2.2 Modeling

The modeling stage is crucial for organizations to meet the accuracy and robustness requirements under Article 15 of the EU AI Act, as well as implement additional risk mitigation measures under Article 9, such as fine-tuning models and incorporating hallucination guardrails (see Table 7). Weights & Biases provides capabilities that assist organizations during this stage of the ML lifecycle, including the following.

AI system tracing

To understand the behaviour of the agent, we trace every hiring decision of the agent and then use the W&B Weave Playground to debug specific hiring decisions and improve the robustness and performance of our prompts (see Figure A & B). Hiring agent decisions can be inspected by technical and non-technical roles with the W&B Weave Playground.

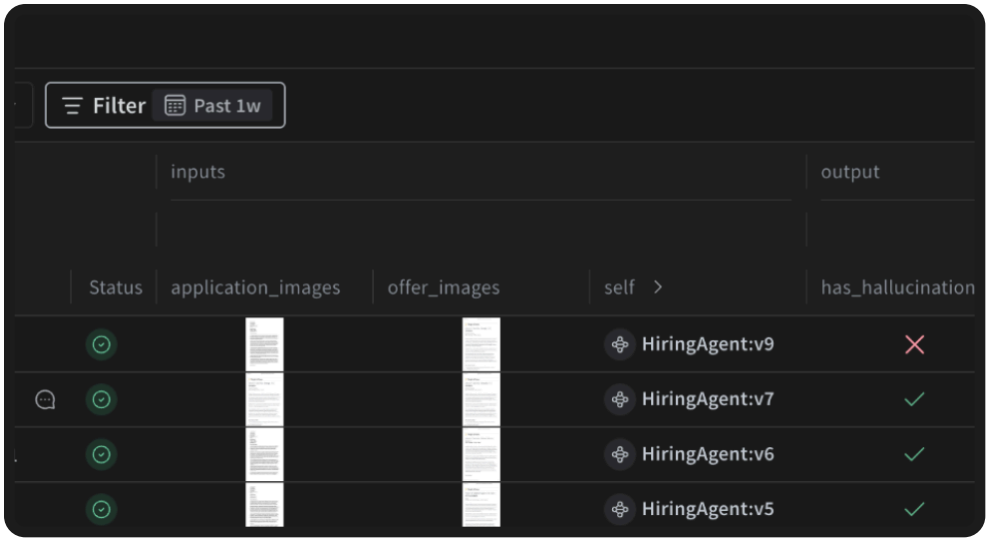


Figure A. Hiring decisions can be debugged and the best performing prompts can be saved

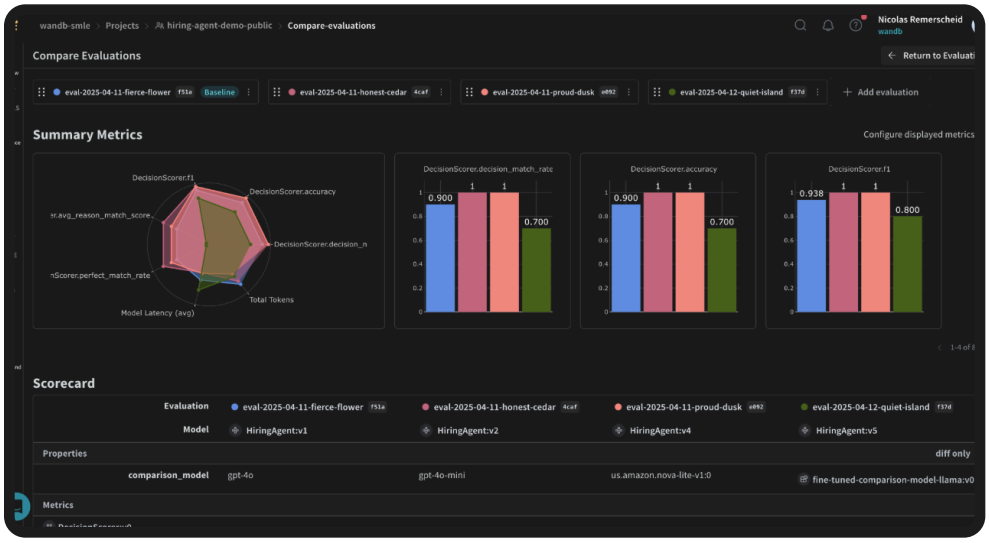


Figure B. Comparing the performance of different hiring agents based on different comparison models

AI system evaluation

We use W&B Weave to evaluate different versions of the hiring agent system across the following performance metrics (see Figures A and B), including decision match (accuracy), hallucination, latency, and costs (see Figure B).

Fine-tuning the comparison model

To mitigate the risk of arbitrary decision making due to hallucinations, we decide fine-tuned the comparison model using W&B Experiments and Sweeps (see Figure C). The figure shows from left to right: training loss over time, parallel coordinate plot visualizing coverage, and parameter importance plot tracking hyperparameter significance.

Figure C. Fine-tuning plots as part of the "Training View" workspace

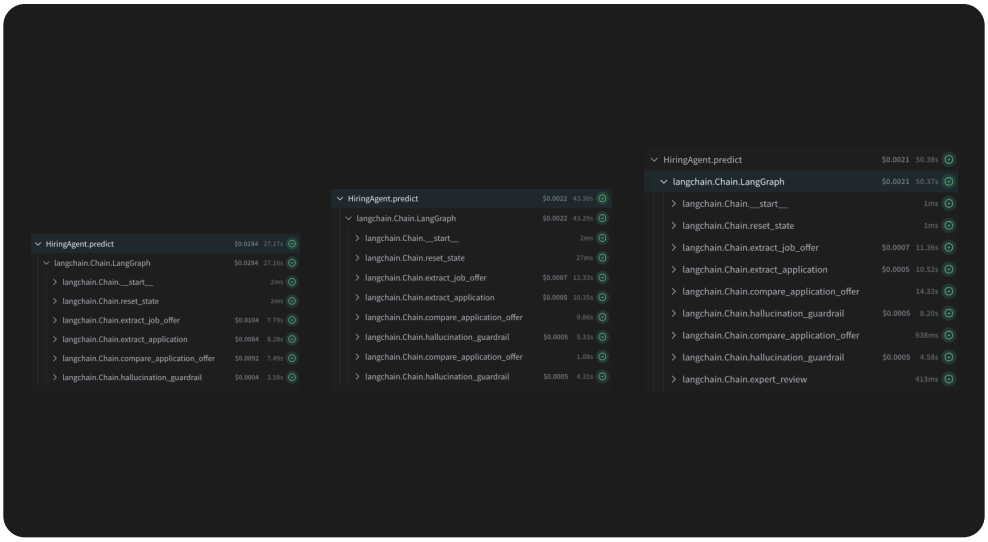


Figure D. Three distinct agentic workflows based on whether the hallucination guardrail detects a hallucination

Guardrails and robustness

To further mitigate hallucination risks, we implemented a guardrail using W&B Weave to verify that the agent's reasoning directly follows from the candidate application or the job description (see Figure D). The images depict three outcomes: a successful comparison passing the guardrail, a self-corrected re-evaluation, and a final failure requiring human intervention.

3.2.3 Deployment and monitoring

Deployment and monitoring are critical to ensuring the AI Hiring Agent's performance, safety, and compliance throughout its lifecycle. Under the EU AI Act, organizations must document conformity assessments (Article 11), manage documentation for downstream actors (Article 13), establish effective human oversight interfaces (Article 14), and ensure traceability (Article 12). We also address additional risk mitigation practices from Article 9, including PII masking and continuous model evaluation (see Table 7). This case study illustrates deployment and monitoring best practices by emphasizing privacy-preserving traceability of hiring decisions and enabling effective human oversight in production.

Development-to-production handover

Before deployment, we compile a central set of documents using W&B Reports. These are the central pieces of documentation that the engineering team, the platform team, and the governance team assess jointly before validating a new version of the hiring agent for production deployment.

Human oversight interface

In order to enable effective human oversight, we built a simple UI that queries information from the W&B Weave API (see Figure A). This UI allows a human expert operator to observe the models performance and behavior and to control the system by manually disregarding or overriding the decision of the system.

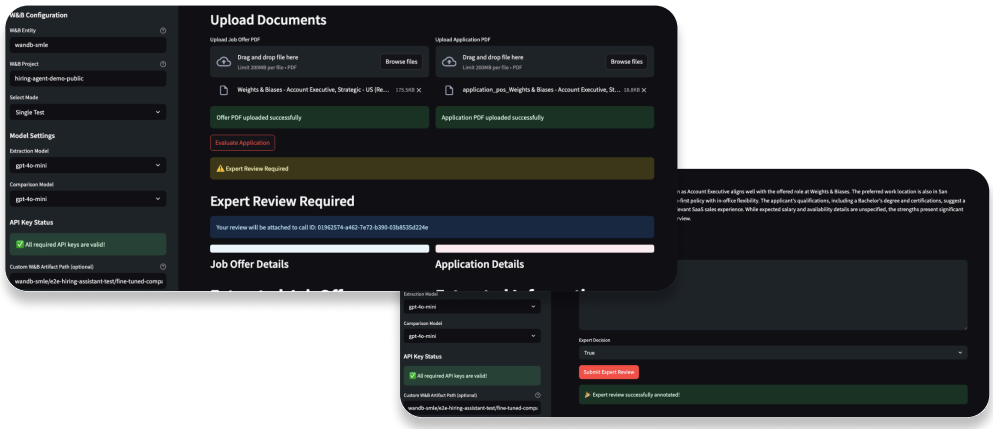


Figure A. A simple application that is used by the human expert operator to override decisions by the hiring agent (shows up if the guardrail failed twice)

Privacy and data protection

To address the potential risk of PII leakage, we use W&B Weave's PII masking function to mask out all PII data that is logged to Weave.

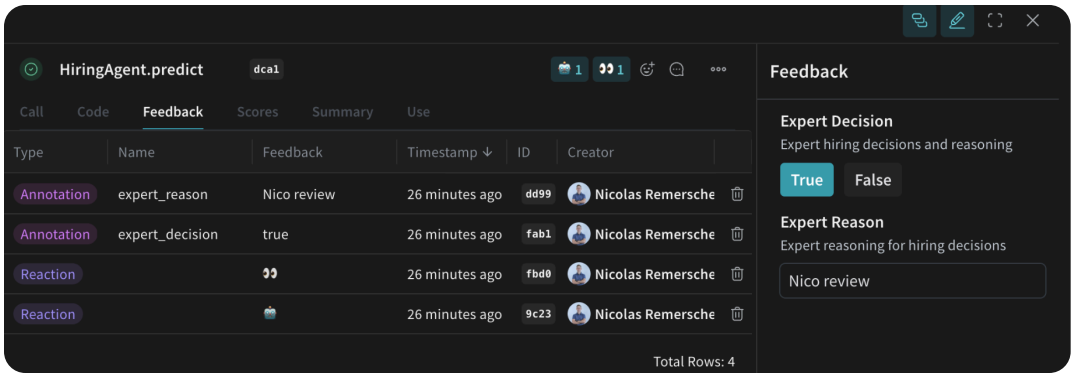


Figure B. Experts can annotate specific production traces from the Weave UI

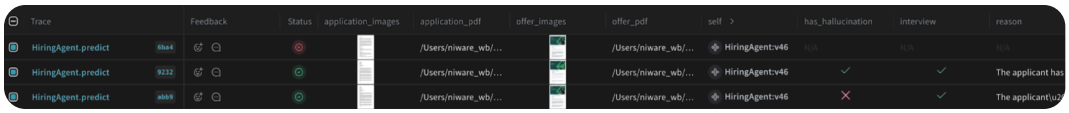


Figure C. Annotated production traces can be added to datasets

Traceability

W&B Weave Traces allow us to monitor the specific decision processes the agent followed in production. This helps us to continuously investigate if the AI system presents a risk (see Figure B and C).

Continuous improvement from production

To monitor the system after deployment, we use Weave Monitors to observe metrics like data drift, model drift, and costs and use annotated production traces to loop back data from production to update our fine-tuning and evaluation dataset in order to make it more representative.

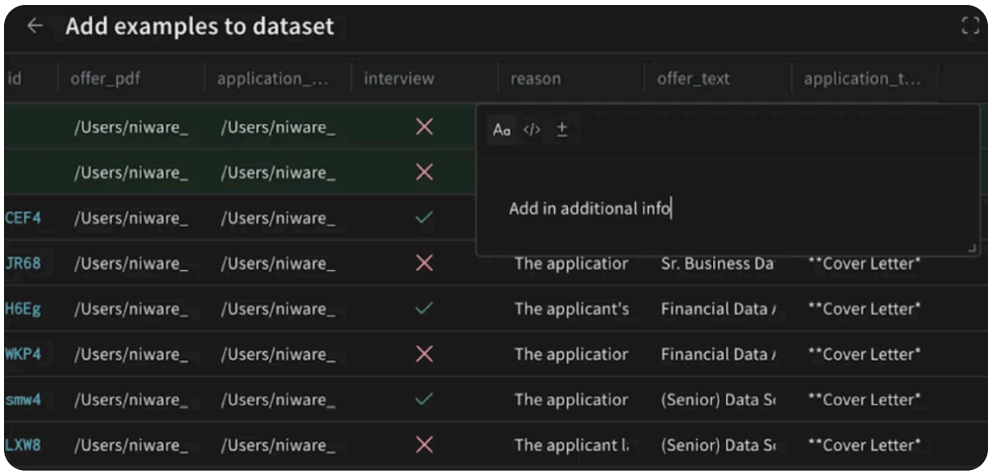


Figure D. Final manual modifications can be done in the dataset editor

3.3 Additional considerations

In previous sections, we demonstrated how Weights & Biases can support enterprises to comply with the EU AI Act:

- Using Weights & Biases for data collection, exploration, evaluation, and lineage tracking, teams can standardize data governance processes, collaborate through seamless hand-offs, and automate reporting.
- With Weights & Biases, teams can fine-tune and evaluate models, and implement guardrails to mitigate risks.
- By leveraging Weights & Biases post-production tools, organizations can establish comprehensive traceability, automated logging, and continuous performance evaluation, real-time monitoring, and human oversight interfaces.

Chapter 3, Section 3 of the EU AI Act outlines additional considerations for providers of high-risk AI systems—including conformity assessment, AI system registering, and other procedural requirements—that are out of scope of this paper.

Conclusion and next steps

As AI adoption accelerates, regulatory compliance becomes not merely an obligation but a strategic imperative. The EU AI Act demands rigorous processes and comprehensive documentation, making the right choice of supporting technology crucial for organizations navigating these requirements. Weights & Biases offers robust, integrated tools that simplify compliance tasks—enabling teams to effectively manage risk, maintain transparency, and deliver trustworthy AI solutions. By leveraging the Weights & Biases AI developer platform, organizations can transform regulatory compliance into an advantage, ensuring continuous innovation while upholding the highest standards of safety and ethical responsibility.

1. **Setup an EU AI Act governance** process to identify your risk landscape, define your regulatory strategy, relevant roles & responsibilities, oversight policies, and evaluate governance across the ML lifecycle. Contact appliedAI Initiative to find out more.
2. **Identify your obligations** by creating an inventory of AI systems in your company and classifying their risk class and role. Contact appliedAI Initiative to find out more.
3. **Standardize your compliance processes** through the Weights & Biases observability and governance platform. Contact us for a demo.
4. **Execute and test** policies for your obligations under the EU AI Act with tools that support collaboration between technical and non-technical stakeholders.
5. **Demonstrate compliance** through technical documentation and conformity assessment when necessary.
6. **Upskill your workforce by** rolling out AI literacy training per role. Contact appliedAI Initiative to find out more.

Check out our end-to-end AI Agent compliance report:

Explore the W&B workspace and generated compliance report with all the technical details and explanations.

About Weights & Biases

Weights & Biases is the AI developer platform powering the AI industry. Over 1,300 organizations worldwide—including AstraZeneca, Canva, NVIDIA, Snowflake, Square, Toyota, and Wayve—and more than 30 foundation model builders, such as OpenAI, Meta, and Cohere, rely on Weights & Biases as their system of record for training and fine-tuning AI models and developing AI agents and applications with confidence.

Our main goal is simple: give developers the best tools to build AI agents, applications, and models. Headquartered in San Francisco with a global presence, Weights & Biases offers a comprehensive platform to help you take AI from idea to production:

- W&B Weave helps you evaluate, monitor, and iterate on AI agents and applications
- W&B Models helps you train, fine-tune, and manage AI models

All of these tools come together in one unified platform, complete with enterprise-level performance, scalability, governance, and security.

About appliedAI Initiative

appliedAI is Europe's largest initiative for the application of cutting-edge trustworthy AI in enterprises. The EU AI Act could put compliance hurdles in the way of businesses. But it also gives us an opportunity to create high-quality AI products and services. appliedAI supports companies to transform regulation and governance into a competitive advantage and not a burden through:

1. EU AI Act governance process: appliedAI supports you to define your compliance strategy, organizational setup, processes, and responsibilities
2. EU AI Act technical implementation: appliedAI supports you in reducing time to compliance by defining a complaint-by-design strategy for the technical implementation of EU AI Act for your AI use cases
3. EU AI Act literacy training: appliedAI helps your business close the AI literacy skill gap with tailored training customized per role and level of AI literacy

appliedAI is your trusted partner that supports implementing the EU AI Act in your company, becoming compliant, and generating a competitive advantage with AI.

Authors

Alexander Machado is the Head of Trustworthy AI CoE and former Head of MLOps Processes. He has a decade of experience in Data Science, Artificial Intelligence, and Data Engineering at appliedAI, the Max Planck Society, and BMW. His work focused on leading, planning, and developing AI solutions from experimentation to production. He has led multiple AI Governance-MLOps working groups, published an MLOps online course, and developed practical frameworks that address the inherent challenges of production systems and compliance with the EU AI Act.

LinkedIn

Nicolas Remerscheid, originally from Geneva, studied Electrical Engineering and later Data Engineering and Analytics at TU Munich. In his final years of study he focused on researching on Privacy-Preserving ML at the Imperial College in London and co-founded one of the largest student AI initiatives in Europe - TUM.ai. After gathering experience at appliedAI and the Swiss Space Force as AI Engineer, he is now working as AI solutions engineer at Weights & Biases with a specific focus on LLMOps.

LinkedIn

Akhil Deo is a senior AI regulatory expert. He has 6 years of experience working on technology and public policy. Akhil was previously a communications specialist at the Future of Life Institute and a junior research fellow at the Observer Research Foundation. At appliedAI, he led a working group with leading European companies on implementing the requirements for high-risk AI systems.

LinkedIn

Chander Matrubhutam is sr. director of product marketing at Weights & Biases where he helps customers and prospects understand the benefits and challenges of implementing AI and the tooling required to observe, iterate, and govern AI agents and applications in real-world use cases.

Linkedin