

OECD AI Transparency Report

Organization: ZAION (FRA) Reporting Period: Q3 2026 Published: October 2, 2026

Section 0 – Your role in the AI system lifecycle

Responding as (please select all that apply):

- Application developer (/provider)
- Deployer

If any of the fields above are selected, please indicate your organization size:

Medium-sized enterprise: 50-249 employees

Section 1 – Risk Identification and Evaluation

1. How does your organization define and/or classify different types of risks related to AI?¹

- Maintain an internal taxonomy, classification system, and/or set of risk area definitions
- Align with specific external frameworks, standards, reports, and/or guidance documents such as taxonomies, classification systems, and/or sets of definitions for risk areas (e.g. the NIST AI Risk Management Framework, the ISO/IEC 42001, the OECD AI Principles, the EU General-Purpose AI Code of Practice, the International AI Safety Report). (please specify): ISO/IEC 42001 for AI management, ISO/IEC 27001 for information security, SOC 2, GDPR and CNIL guidance, and the EU AI Act. We maintain a documented AI risk taxonomy with risks held in a central register. AI governance is integrated with our existing information security and privacy programmes rather than run separately, so the three share governance, policy hierarchy and audit cycles.

How does your organization define thresholds at which severe risks, unless adequately mitigated, may be considered "unreasonable"? Please describe the framework(s) most relevant to your organization's role and context. If your organisation does not use formal thresholds, you may describe instead the policies, decision criteria, or governance processes you apply when assessing severe AI risks. You may also describe: • Specific examples of contexts in which your models or systems could, in your assessment, pose unreasonable or intolerable risk; • How mitigations are graduated by capability or propensity level; • How your organisation considers the specificities of different AI technologies, such as generative AI and agentic AI, including any specific risk assessment or governance mechanisms applied to agentic AI capabilities.

Zaion is a provider of conversational AI for customer relations (Voice and Chat AI) for clients in insurance, banking, health, assistance, public services, utilities, transport and telecommunications. We

do not train foundation or general-purpose models. We develop proprietary speech and language components and integrate third-party foundation models under contractual and technical controls. Our thresholds are harm-severity-based rather than capability-based: we assess impact, reversibility and scale in the specific deployment context (consistent with NIST AI RMF), rather than against model capability thresholds, which apply to frontier model developers.

Contexts we assess as unreasonable absent mitigation: (i) a caller unable to reach a human when circumstances require it, including distress or a safeguarding concern; (ii) automated handling producing legal or similarly significant effects on a caller without human review; (iii) deployment without disclosure that the caller is speaking with an automated system; (iv) systematically degraded recognition for an identifiable caller group, producing unequal access to a service.

Graduation of mitigations: conversational scope is the primary control. An intent is admitted to automation only where accuracy and escalation behaviour meet acceptance criteria; where they cannot be met reliably the intent is withdrawn from automated scope. Confidence thresholds trigger escalation, so a recognition or comprehension failure degrades into a human conversation rather than a failed service interaction.

Generative and agentic specificity: generative responses are grounded in client-approved knowledge sources rather than open-ended generation. Also, agentic capabilities are governed by an action allowlist based on client specification and a declared-tools-only policy to block faulty actions.

2. Both before and after deployment, what practices does your organization use to identify and evaluate vulnerabilities, incidents, emerging risks and misuse throughout the AI system lifecycle?

- Adversarial testing to detect technical vulnerabilities
- Internal benchmarks, standard metrics (e.g., frequency, severity scores)
- Structured techniques to explore the behavior or performance of a model and/or application, such as model elicitation or bottleneck assessments
- Qualitative assessments (e.g., expert judgment, narrative scoring)
- Impact assessments such as personal data protection assessment and fundamental rights impact assessments
- Internal model, system, or application-independent research initiatives
- Incident reports, including reports shared by other organizations
- Automated monitoring / remote status reporting (e.g., telemetry pipelines, automated alerts when risk thresholds are breached)
- Specific risk identification practices for agentic AI capabilities, e.g., autonomous action-taking, tool use, multi-step planning (please specify): Bounded agent steps (recursion limit), action allowlist, declared-tools-only execution, tool-error capture
- Assessment of risks arising from interactions among multiple AI agents
- Multilingual and cross-cultural evaluation, including collaboration with local ecosystems (e.g., national AI institutes, language-specific evaluation consortia) for under-represented languages and cultural contexts (please specify): Zaion operates multilingual conversational AI systems using both its own proprietary and third-party models. We frequently evaluate and measure performance for different populations and run gender-stratified evaluation of our speech models against our fairness target, and expert linguists review

model and voice quality. We document the limitation openly to tackle them based on our R&D approach. All the evaluations are run by our internal teams, which are both multidisciplinary and multicultural

How does your organization monitor the interactions of AI agents under your control with other agents / assess risk from multiple agents interacting?

Bounded agent steps through a recursion limit; an allowlist of permitted actions; execution restricted to declared tools only; capture of tool errors; and loop detection applied to both input and output.

Multi-agent interaction: the gateway agent and its sub-agents interact with the platform orchestrator and with tools. Those interactions are bounded by loop detection across the conversation history, a per-call request cap and per-call and per-conversation token budgets. Every agent step is recorded in the trace returned to the orchestrator, so the deploying organisation can see exactly what was done.

Please briefly describe your organization's decision-making process for deploying an AI model or system including: who has decision authority; what evidence and criteria are required (e.g., performance, safety/security, robustness, fairness, privacy, compliance, independent external testing, evaluation, validation and verification (TEVV) or assurance); contextual caveats or limitations required in interpreting risk evaluation metric results; what approvals/sign-offs are needed; and what safeguards (e.g., mitigations, monitoring, or phased release) must be in place to move into deployment.

Zaion operates multilingual conversational AI systems using both its own proprietary and third-party models. We frequently evaluate and measure performance for different populations and run gender-stratified evaluation of our speech models against our fairness target, and expert linguists review model and voice quality. We document the limitation openly to tackle them based on our R&D approach. All the evaluations are run by our internal teams, which are both multidisciplinary and multicultural

Following deployment, how do you use incident reports — including those shared by other organizations — to identify emerging issues, update risk assessments, and trigger corrective actions.

Incidents are captured automatically rather than waiting to be reported. Every response carries a structured trace recording the error type and description, which provider and model were used, and any warnings, correlated by request identifier with our metrics. Recurrent issues are analysed, a fix and a regression test are added together, and thresholds or configuration are updated. Provider-side incidents such as outages or content-filter events are handled automatically by fallback and circuit breaking before they reach the caller.

At platform level we monitor accuracy, drift and output quality, fairness degradation against threshold, and security signals including adversarial attacks, data poisoning and unusual inference patterns.

Model performance - accuracy, fairness, explainability and drift - is reviewed quarterly. Incident logs are analysed monthly. Impact assessment coverage is reported quarterly.

Escalation clocks are defined. An incident or complaint triggers a fresh impact assessment. Where any monitored metric falls below threshold, the owner has specific period to escalate with a remediation plan.

3. Does your organization have mechanisms to receive and/or share reports of risks, incidents or vulnerabilities by/with third parties?

Yes, both receiving and sharing

If yes, please indicate which stakeholders reporting mechanisms may be accessible to, recognising specifics will depend on the nature of the report and risk level. For each, indicate whether access is on a standing basis (Always), activated based on context (Ad hoc/as relevant), or not currently accessible.

	Always	Ad hoc / as relevant	Not currently accessible
Relevant employees (security, legal/compliance, etc.)	☐		
Public authorities (e.g., public reporting offices for security incidents or data breaches)		☐	
Supervisory authorities (e.g., data protection authorities)		☐	
Industry partners or industry fora		☐	
Users (end users, enterprise customers, developers using APIs)		☐	
Business customers / deploying organizations (if different from “users”)		☐	
Supply-chain partners (cloud, data providers, etc.)			☐
Academia / researchers			
Independent auditors		☐	
Standards bodies / multi-stakeholder initiatives (OECD.AI, ISO/IEC, IEEE, etc.)		☐	
Specific civil society organizations			
Media / general public (people or communities indirectly impacted)			
Other			

Please describe how your organisation determines when and how reporting mechanisms are made accessible to different stakeholders (e.g. based on legal or regulatory requirements, internal policies, or risk level). Please also provide public URLs or details on how to submit reports. If none, write “Not applicable”

Access depends on legal obligation, contractual commitment and risk level.

For staff: a confidential, explicitly retaliation-free process for raising concerns about AI misuse, unethical outputs, bias, safety, security vulnerabilities or policy breaches. Reports go to a dedicated mailbox, directly to the AI Lead or CISO and all are logged in the AI management incident register. Reporter identity is protected as far as the law allows and retaliation is itself a disciplinary offence.

For clients: enterprise customers have standing account and support channels with SLA-backed escalation, and in practice the diagnostic trace returned with every request makes reporting a problem straightforward - they can see which model answered, what it did and what went wrong. Product

documentation must carry a dedicated route for AI-related questions and an escalation path for failures, unexpected outputs or bias concerns. Customers, partners and suppliers are also served through our Trust Centre.

Public URLs: <https://developers.zaion.ai/docs/platform/security/overview>

How does your organization share relevant information as appropriate?

- Information-sharing partnerships (with confidentiality agreements as appropriate)
- Reporting to regulatory authorities as relevant

Please describe what type of information is shared.

Policy updates to all employees annually or on revision. Risk and incident reports to compliance, executives and regulators where required. System performance reports to leadership and engineering twice a year. Management review outcomes to the executive team annually. Supplier compliance updates twice a year through the Trust Centre. Customer transparency and ethics reports annually through the Trust Centre.

Incident content shared with clients covers what happened, its scope, which data categories were affected and what we did. Personal data breaches go to CNIL within 72 hours and to affected individuals without undue delay where their risk is high. Serious AI incidents go to the competent national authority without undue delay.

4. If your organization has programs to incentivize risk(s) disclosure, incidents and systemic vulnerabilities, please share what type of programs you use?

- Internal incentives and culture

Please provide public URLs and/or details on your incentive program(s). If none, write "Not applicable."

Zaion does not run a paid bug bounty or a formal external disclosure programme with safe harbour. What we do run is an internal confidential, explicitly retaliation-free channel for AI ethics, bias, safety and security concerns, open to every employee and contractor. Also, Zaion maintain an engineering culture of writing findings down - findings reports, risks flagged in service documentation, peer review.

5. Does your organization use or contribute to the development of international technical standards, best practices or tools for the identification, assessment, and evaluation of risks?

Uses

If contributes or uses, please select all that apply:

- ISO/IEC standards (e.g., ISO/IEC 27001, ISO/IEC 42001, ISO/IEC 23894, ISO/IEC 42005)
- Industry-led best practices or codes of conduct

6. Does your organization collaborate with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks that could have broad and potentially cascading socio-economic impacts? Please indicate which stakeholders you collaborate with, and whether collaboration is on a standing basis (Always-on), activated based on context (Ad hoc/as relevant), or there is no current collaboration.

	Always-on	Ad hoc / as relevant	No current collaboration
Relevant employees (security, legal/compliance, etc.)	☐		
Public authorities / regulators / AI safety institutes / other public sector bodies and inter-governmental coordination structures		☐	
Industry partners or industry fora		☐	
Users (end users, enterprise customers, developers using APIs)		☐	
Business customers / deploying organizations (if different from “users”)		☐	
Supply-chain partners (cloud, data providers, etc.)			☐
Academia / researchers		☐	
Independent auditors		☐	
Standards bodies / multi-stakeholder initiatives (OECD.AI; ISO/IEC; IEEE; AIID etc.)		☐	
Independent civil society organizations (e.g., for evaluations, audits, or policy review)			
Media / general public (people or communities indirectly impacted)			
Other(s) (please add any other relevant stakeholder groups):			
Not applicable			

Please describe examples of how your organization collaborates. Examples may include: participating in multi-stakeholder or industry initiatives; partnering with academia, civil society, or the public sector; establishing internal or joint governance structures with external experts or in partnership with other efforts to share knowledge and jointly address risks.

Zaion's collaboration takes place mainly through its existing working relationships rather than formal external initiatives. We work with the organisations deploying our systems to define the scope of each use case and the safeguards appropriate to it, and with our cloud and model suppliers on the security, privacy and reliability of the components we integrate. Internally, our technical, security, compliance and

product functions address AI risk together within a shared governance framework. We also follow the work of standards bodies, regulators and industry groups relevant to our sector.

Any further comments and implementation documentation.

No answer provided

Section 2 – Risk Management and Information Security

7. Across the AI system lifecycle, what steps does your organization take to manage risks and vulnerabilities? Please indicate which specific practices, tools or metrics your organization uses at each lifecycle stage. You may refer to tools or metrics present in the OECD.AI Catalogue of Tools and Metrics or suggest additions. Respondents may complete only the lifecycle stages relevant to their organisation's role. Stages that do not apply may be left blank.

Plan and design

Collect and process data

Build, integrate and interpret model

Verify and validate

Deploy

Operate and monitor

Please provide a consolidated description of your organisation's approach to risk management across the AI system lifecycle, including how the practices identified above operate together and any cross-cutting governance, review or escalation processes. Please also indicate how often these practices are reviewed and what types of events trigger reviews and updates. Where relevant, please distinguish between different types of AI systems (e.g. generative or agentic AI).

Risk management is built into the services as layered automatic controls rather than sitting only in documents :

- Preventive: allowlists of models, actions and tools; input validation; personal-data masking; token caps.

- Detective: typed traces, metrics, loop detection, an alignment quality signal.

- Corrective: fallback providers, circuit breaker, retry without tools, fail-closed bleeping, and a safe default response that hands an explicit error state back to the orchestrator rather than improvising.

The three system types are managed differently.

- Generative: risk comes mainly from third-party model behaviour, managed by output schema validation, content-filter handling and provider fallback.

- Agentic: bounded steps, action allowlist, loop detection.

- The discriminative personal-data model: threshold calibration on a gold set and fail-closed design.

Above that, the six-gate lifecycle governs progression, each gate with a named condition and a named approver, and escalation running from the AI Lead to the CISO to top management.

Cadence : Every release re-runs tests and metrics. Every new model version triggers gold-set re-evaluation. Every provider change updates the allowlist. Incidents and findings trigger a fix plus a regression test.

- Quarterly: model performance, data quality, audit findings, objectives, risk treatment, impact assessment coverage, and an operational status report.
- Monthly: incident analysis.
- Twice a year: supplier compliance, performance reporting to leadership, data quality review.
- Annually: internal audit across seven areas, refreshed impact assessments, gap assessment, training completion, management review, and review of every policy. Registers and provenance records are kept at least three years or for the operational life of the system.

Please describe any measures used to ensure data quality and fairness across the data collection and curation stage (e.g., dataset audits, filtering training data to remove harmful content or personal information, fine-tuning on curated data, use of moderation tools, assessment of limitations to model generalisability).

Quality dimensions are defined and measured for every dataset: accuracy, completeness, consistency, timeliness, representativeness, and bias and fairness. Thresholds: completeness at or above 95%, operational data refreshed within 30 days, fairness variance below 10% across demographic groups. Bias and representativeness results must be reviewed before any production deployment, and live systems get a further data quality review at least every six months. For the retrieval-based chatbot we also track knowledge-base freshness, accuracy and completeness, and limit irrelevant context using a similarity threshold and metadata filters.

We also run gender-stratified evaluation of our speech models against our fairness target, and expert linguists review model and voice quality.

Known limitations, recorded rather than hidden.

8. What security measures are used in testing AI systems?

- Isolated and secure test environments
- Network isolation from external access
- Locally hosted testing environments
- Restrictions on outbound Internet or API access, including a list of permitted domains or APIs
- Isolation of staging environments (from production / operational environments)
- Controlled environments for red-teaming and adversarial testing
- Containment measures for prompt-injection and jailbreak attempts
- Controlled security vulnerability testing
- Rate limits and resource caps
- Human-in-the-loop controls for higher-risk test cases
- Access controls, including roles-based permissions
- Audit logging and monitoring
- Use of “AI-safe” test-data, as appropriate
- Use of synthetic, anonymized, or dummy data, as appropriate

8.A. Please indicate, if relevant, the team or role responsible for overseeing secure testing and whether this responsibility is shared across multiple teams and/or functions.

Zaion Lab - our AI research team, and our infra and security team oversees secure testing of the AI services. Responsibility is shared with Infrastructure and Security, who hold custody of production data, environments and secrets, and with the platform and product teams.

Separation of duties applies: people who develop a system cannot be its sole approvers or auditors, and internal auditors cannot audit areas they are operationally responsible for. Where team size makes overlap unavoidable, compensating controls such as peer review and management sign-off are documented.

9. How does your organization protect intellectual property, including copyright-protected content?

9.A. Safeguards that promote lawful data use and transparency in AI training:

- Maintain a copyright policy that addresses data sourcing, content provenance, safeguards, and organizational roles and responsibilities
- Apply provenance controls to guard against training data being sourced from websites or repositories that have been determined to engage in large-scale piracy
- Refrain from the unlawful circumvention of technological protection measures in data sourcing
- Conduct data cleaning, qualification, and/or tokenization activities as part of a transformative data preparation process
- Consult and/or retain an intellectual property attorney to protect, enforce, and comply with, intellectual property rights.

9.B. Safeguards that reduce the risk of dissemination of outputs that may infringe copyright:

- Restricting outputs that raise ethical, human rights, or legal concerns—such as non-consensual intimate imagery, unauthorized deepfakes, or outputs that are substantially similar to copyrighted works.
- Filtering results using keyword or image-based restrictions to reduce risks of misuse

10. How do you protect privacy and prevent your systems from disclosing confidential or sensitive data across the AI lifecycle?

- Make information about data privacy, practices, and controls accessible to a diverse set of stakeholders
- Use security controls, including encryption at rest and in transit
- Provide organizational customers with controls to configure and implement privacy choices (e.g., data retention, logging, access, or processing settings)
- Apply Privacy-Enhancing Technologies (PETs) where applicable (e.g. differential privacy, federated learning, synthetic data generation, homomorphic encryption)

Describe key privacy measures and data-protection standards applied.

Minimisation and pseudonymisation are the centre of the design, not an add-on.

The anonymisation service removes personal data from call transcripts and returns bleep ranges for the audio. Before any prompt, conversation history or tool output reaches a language model, the gateway

can replace emails, phone numbers, bank details, social security numbers and URLs with reversible placeholders, restoring them only in the response that goes back to the customer. Sensitive variables can be hidden from the model entirely. Detection is deliberately tuned for recall: a word is masked as soon as it reaches a 10% probability of being personal data, and if transcript-to-audio alignment quality falls below 80% the service bleeps the whole utterance rather than risk a partial leak.

Neither shared service persists conversation content. Processing is in-memory, agent memory is cleared after each request, and the cache holds counters and identifiers only, with expiry.

Data is encrypted with AES-256 at rest and TLS in transit, with least-privilege access, per-tenant segregation and secrets held in a dedicated vault scoped per workspace. Personal data in development and test environments must be anonymised or pseudonymised unless the DPO formally approves otherwise. A data protection impact assessment is required before deploying any system that processes personal data.

Providers are treated as sub-processors with documented records confirming that no Zaion or client data trains their models. Client data remains the client's property, and its reuse for training or for other clients is contractually excluded. Customers configure masking and processing choices per workspace and per request.

Where the data sits. Production workloads run on cloud infrastructure in EU regions, with object storage encrypted server-side and reachable only through scoped roles. Managed databases run on private cloud in France - Paris, Marseille and Lyon - encrypted at rest and reachable only through private endpoints. Some legacy infrastructure remains elsewhere in the EU and is being consolidated. Cloud suppliers must hold ISO 27001 and SOC 2, confirm EU data residency, and have a data processing agreement in place.

11. How does your organization assess cybersecurity risks and implement policies to enhance the cybersecurity of advanced AI systems across the AI system lifecycle? Respondents may complete the checkbox subsections in 11.A–11.D or describe their overall approach under 11.E.

11.A. Governance, policy and assessments

- Conduct regular cybersecurity risk and vulnerability assessments tailored to AI models or systems
- Perform threat modeling that includes AI-specific risks (e.g., data poisoning, prompt injection)
- Maintain or update security policies aligned with recognized standards (e.g., NIST CSF, ISO/IEC 27001, SOC 2, CIS Critical Security Controls), and review them regularly in response to emerging risks or regulatory guidance
- Maintain vulnerability management processes (scanning, channels for external reporting, patching, remediation)
- Implement incident response, disaster recovery, or business continuity plans, including escalation procedures for high-severity incidents
- Train staff on cybersecurity best practices and AI-specific threats, including how to recognize and respond to misuse or unintended outcomes of advanced AI systems
- Integrate cybersecurity AI risks into broader cybersecurity risk management activities
- Regularly review security measures and update them in response to emerging risks, incidents, or major releases

11.B. Technical and operational controls

- Apply multi-factor authentication (MFA) and role-based access controls for advanced AI systems
- Encrypt AI assets (datasets, models, configurations) in storage and in transit as appropriate
- Monitor system activity and log access to detect anomalies or misuse
- Conduct secure AI development and/or deployment practices
- Conduct security testing focused on AI-specific risks (e.g., red-teaming, adversarial testing, stress tests)

11.C. Protection of sensitive or high value AI assets

- Limit access to proprietary AI assets based on role or need-to-know
- Store model weights, datasets, and algorithms in secure, access-controlled environments
- Track and log access to sensitive AI resources
- Require NDAs or confidentiality agreements for staff and contractors with access

11.D. Insider risk protection detections

- Conduct pre-employment screening or role-sensitive vetting
- Monitor for unauthorized internal access or suspicious activity
- Provide staff training on cybersecurity best practices and AI-specific threats
- Implement insider-risk escalation or response procedures

11.E. Please describe the main security measures your organization applies and indicate which are specific to particularly sensitive AI assets, where relevant.

- AES-256 at rest and TLS in transit. Least-privilege access and per-tenant segregation. Secrets in a dedicated vault with workload authentication and per-workspace isolation. API-key authentication separating user and administrative scopes. Rate limiting, per-call request caps and token budgets, which protect against cost and abuse attacks as much as against load. Input validation through strict schemas, transcript and timestamp validation, and bounded conversation history. Allowlists for models and for agent actions. Provider content filtering. Circuit breaker and fallbacks for availability. Containers built from minimal images, run non-root and pinned by digest.

- Model weights are kept out of source control and out of container images; provider credentials never appear in code; personal data is masked before anything reaches a third-party model. Model artefacts and training data sit in encrypted storage reachable only through scoped roles, with versioning and audit-log immutability. All AI assets carry a security classification, and the production engines are classified Confidential.

The AI-specific threat surface is treated as a security matter rather than a quality one. Threat modelling covers data poisoning, prompt injection and agentic security. In a voice product the caller's own speech is untrusted input reaching a model, which is exactly why bounded scope, allowlists, loop detection and refuse-and-escalate behaviour function as security controls. Production monitoring covers adversarial attacks, data poisoning and unusual inference patterns.

Some responsibilities are contractually delegated: physical security of compute and storage sits with our cloud providers, who must evidence ISO 27001, SOC 2 and EU data residency and are assessed annually.

12. What is your organization's vulnerability management process? Does your organization take actions to address identified risks and vulnerabilities, including in collaboration with other stakeholders?

- Internal vulnerability disclosure receipt and tracking processes
- Investigation of reported or identified risks and vulnerabilities
- Development of patches or other mitigations for confirmed risks or vulnerabilities
- Testing of patches or other mitigations
- Scanning for and mitigation of similar risks or vulnerabilities
- Reporting to regulators / authorities when relevant

13. What post-deployment practices does your organization use to monitor and respond to vulnerabilities, incidents, emerging risks and misuse of advanced AI systems?

- Conduct post-deployment monitoring
- Perform root-cause analysis and apply lessons learned to improve future development and/or deployment practices
- Monitoring and emergency controls specific to agentic AI systems (e.g., ability to interrupt autonomous action sequences, override tool-use permissions, or revoke delegated authority)

14. Do you develop or deploy agentic AI systems (systems that can autonomously plan/act or use tools)?⁴

Yes

14.A. If yes, please indicate which categories of controls your organization applies. Please select all.

- Constraints on the agent's action space (e.g., tool-use permissions, environment sandboxing, allowlists/blocklists of resources, scope limits per task)
- Human oversight or interruptibility mechanisms (approval requirements for sensitive actions, escalation paths, interruption interfaces)
- Monitoring (anomaly detection, multi-agent interaction monitoring)
- Logging of agent actions

Please provide any further comments and implementation documentation.

Constraints on the action space: an allowlist of permitted actions declared per workspace by the deploying customer; only declared tools may be executed; scope limits per task; a recursion limit capping the agent reasoning or tool steps.

Human oversight and interruptibility: The deploying organisation defines which actions the agent may take at all. Loop detection and the step limit interrupt an agent that is not converging. A failing model or provider is quarantined by the circuit breaker after three. When any of these fires, the agent returns a safe default response with an explicit error state to the orchestrator, which hands control back rather than continuing autonomously.

Monitoring: Loop detection over conversation history; a per-call request cap and per-call and per-conversation token budgets; anomaly and error metrics; monitoring of interaction between the agent, its sub-agents and the orchestrator.

Logging: Every agent step, tool call and its arguments and results are recorded in the trace returned to the deploying organisation, so the actions taken are visible to the party accountable for them rather than only to us.

Section 3 – Transparency Reporting on Advanced AI Systems

15. Does your organization publish public reports and/or technical documentation related to the capabilities, limitations, and domains of appropriate and inappropriate use of advanced AI systems? Are these clear and understandable by non-technical audiences?

No

16. How does your organization share information with a diverse set of stakeholders (other organizations, governments, civil society and academia, etc.) regarding the outcome of evaluations of risks and impacts related to an advanced AI system?

- Not applicable

Describe how evaluation results are shared and with which stakeholders (e.g. other organizations, governments, civil society and academia, etc). You may also indicate whether non-technical or plain-language summaries are provided for broader audiences.

Evaluation and performance results are shared primarily with the organisations deploying our systems, under the confidentiality terms of our client agreements, through service documentation and regular service reviews. Results are reported to internal leadership as part of our management review cycle, and to regulators where required. Material is adapted to its audience, with plain-language summaries provided for non-technical readers where appropriate.

17. Does your organization publish public privacy policies addressing the use of personal data, user prompts, and/or the outputs of advanced AI systems?

No

18. How does your organization provide information about the sources of data used for the training of advanced AI systems, as appropriate, including information related to the sourcing of data annotation and enrichment?

- Protected information sharing
- Regulatory reporting as relevant

Please provide links to any publicly available information.

Information on data sources is provided to clients and downstream partners rather than published openly. We maintain a structured register recording, for every dataset, its source and provenance, rights and licences, privacy check, type and category, what it is used for, refresh cadence, labelling and preparation method, documented bias and risk notes, retention reference, lawful basis and owner. Training data for third-party speech and language components is governed by those providers' own documentation, with our agreements confirming that our data does not train their models.

19. Does your organization demonstrate transparency related to advanced AI systems through any other methods?

Yes

If yes, describe additional method(s), including where relevant transparency measures related to AI agents

Yes, for the users of our conversational IA solutions (Voice AI Agent, Chat AI Agent), the transparency measure with the widest reach is disclosure at the moment of interaction, mandated per product. AI Agent gives the user a clear notice at the start of the call. Agent-Assist is visible in the agent interface

20. Does your organization publish statistical insights derived from anonymised, aggregated usage data to advance understanding of AI adoption and its economic impacts in areas including occupations and tasks by region, sector and country? To this end, do you contribute to international cooperation efforts?

No

Please estimate the total staff time and annual expenses associated with conducting compliance activities and include any further comments and implementation documentation (optional).

No answer provided

Section 4 – Organizational Governance, Incident Management, and Transparency

21. How has AI risk management been embedded in your organization's governance framework?

- Executive oversight
- Senior management oversight

Please describe how AI risk management is integrated into governance structures, including: which roles or bodies have advisory versus decision-making authority; which roles or bodies have final accountability for AI risk-related decisions; and how often policies are reviewed or updated (e.g., following major incidents, regulatory changes, or material updates to systems).

AI risk management is embedded through an ISO/IEC 42001-aligned AI management system:

- Accountability. The CISO, who is also Head of Infrastructure and Security, is accountable for AI governance and for every AI policy; the AI Lead is responsible for implementation. Top management

holds final authority: it approves major objectives and resourcing, reviews risk reports and audit findings, formally accepts or rejects residual risk that cannot be mitigated, and approves major external AI communications.

Zaion Lab documents systems, datasets and model cards. Infrastructure and Security hold custody of production data, environments and secrets. Service owners flag and escalate open risks for management sign-off. Separation of duties is mandatory throughout.

Our governance is deliberately lightweight for our size: named ownership of every AI decision and risk, AI risk and performance reported directly to the CISO and top management, and leadership, technical teams and compliance working inside one shared framework rather than through layers of committees.

22. Are relevant staff trained on your organization's governance policies and risk management practices?

Yes

If yes, describe training methods (e.g., onboarding modules, regular sessions, and confidential channels to report concerns).

Training is role-based, at appointment or onboarding and annually thereafter. The AI Lead takes ISO 42001 foundation training plus AI ethics and governance. The CISO takes ISO 27001 implementer training plus AI security awareness. Product and engineering take secure development, AI ethics and responsible AI principles. The DPO and compliance take data protection, EU AI Act awareness and ISO 42001 foundation. Every employee takes AI ethics awareness, acceptable use of AI tools, and the basics of bias and fairness. Completion is tracked by HR and reported to the CISO. Delivered courses include ISO 42001 foundation, an internal AI ethics course and a fairness testing workshop. For engineering teams this is reinforced operationally: onboarding through detailed service documentation - configuration guides, test and staging acceptance plans, findings reports - plus code review and knowledge sharing between research and platform teams. AI management policies are published internally for all relevant staff.

23. Are steps taken to address reported incidents documented and maintained internally?

Yes

If yes, describe how incident records are kept and used for follow-up (e.g., incident logs, root-cause analysis reports, internal tracking systems). You may indicate whether and how you define escalation thresholds for when incidents must be surfaced to senior leadership or external parties.

Incidents are recorded and reviewed as part of the management review. Technical incident data is captured automatically alongside it: structured traces recording error type, description, provider and model, and warnings; logs carrying request identifiers and session context; and metrics. Fixes are recorded in version history together with a regression test, and analyses are written up as findings documents. Records are kept at least three years or for the operational life of the system.

24. Does your organization share research and best practices on addressing or managing risk?

No

25. Does your organization use international technical standards or best practices for AI risk management and governance policies?

Yes

If yes, please indicate which standards or frameworks apply:

- ISO/IEC standards (e.g., ISO/IEC 27001, ISO/IEC 42001)
- Industry-led best practices or codes of conduct

If yes, describe which international technical standards or best practices your organization uses.

ISO/IEC 42001 is the framework for our AI management system, with a full statement of applicability covering every clause and control, each assessed and tracked individually. Our data management policy and register are aligned to its data acquisition, quality, provenance and preparation controls, with model cards and impact assessments per system. It is integrated with our ISO/IEC 27001 information security management system, sharing governance, policy hierarchy and audit cycles.

Any further comments and implementation documentation.

No answer provided

Section 5 – Content Authentication & Provenance Mechanisms

26. What mechanisms, if any, does your organization put in place to allow end-users, where possible and appropriate, to know when they are interacting with an advanced AI system developed by your organization?

- User disclosure (e.g., text notice, verbal statement, or system label)
- Distinct branding or labeling for AI-generated content (e.g., “AI-generated” tag, metadata label)
- Documentation or FAQs explaining when and how AI is used in the system

Provide any additional details or context on how or when users are informed they are interacting with advanced AI systems (e.g. if such mechanisms are planned, in testing, or applied only to certain products).

Disclosure is mandated per product. AI Agent (Voice and Chat) gives the end-user a clear notice at the start of the interaction. Agent-Assist is visible in the agent interface. BotCenter and Zaion Assistant carries a notice in the admin interface. Wording held in our documentation includes phrasing along the lines of 'you are being assisted by an automated AI system' and 'this call is handled by an automated voice AI'.

Disclosure must be visible and prominent rather than buried in terms and conditions, in plain language suited to the audience, and delivered at the point of interaction. Documentation must also explain how to override or stop the AI, how to escalate to a human, and where human review is expected - with

particular weight on decisions that significantly affect an individual. Failing to disclose AI interaction is an enforceable policy breach.

27. Does your organization use content provenance detection, labeling or watermarking mechanisms that enable users to identify content generated by advanced AI systems?

No

Any further comments and implementation documentation.

No answer provided

Section 6 – Research & Investment to Advance AI Safety & Mitigate Societal Risks

28. Does your organization advance research and investment in the following areas?

- Safety
- Data quality and representativeness
- System performance and outcomes (e.g. accuracy, consistency)
- Explainability or interpretability
- Transparency
- Robustness or reliability

If yes, please describe key research or investment initiatives in the areas selected above (e.g., dedicated safety teams, funding of research programs, contributions to open-source safety tools, publishing research, partnerships and collaborative research initiatives). You may also indicate whether related outputs are publicly shared, shared with partners or within consortia or not publicly disclosed.

In-house research on privacy-preserving speech analytics. We built a multilingual personal-data detection using model-assisted annotation at scale with recall-oriented calibration, entity gap-filling so partial matches cannot leak half an identifier, quantisation-aware training and optimised inference, and timestamp alignment for audio bleeping with fail-closed behaviour when alignment quality drops. We are also work with Academic actors through Thesis programs to advance system performance and outcomes and publish open-source models for the French IA community.

29. Does your organization pursue research or investment related to socio-economic or environmental considerations in the development and deployment of AI?

- Environmental considerations (e.g., energy use, computational efficiency, resource use)

If yes, please provide examples of relevant research or investment initiatives for the areas selected above.

We focus on low footprint environmental and computational efficiency for our proprietary models. We choose CPU-first models where they are adequate rather than defaulting to accelerators. Compute

footprint is monitored as part of resource management, with autoscaling and storage lifecycle policies. We defend a hybrid approach where small models are used when needed and when sufficient, both for classification tasks (NLU) or the use of SLM when no reasoning power is needed.

Any further comments and implementation documentation.

No answer provided

Section 7 – Advancing Human and Global Interests

30. Does your organization support any digital literacy, education or training initiatives to improve user awareness and/or help people understand the nature, capabilities and limitations of advanced AI systems?

No

31. Does your organization pursue AI projects that aim to deliver societal or public benefits, including through partnerships with civil society and community groups?

No

Any further comments and implementation documentation.

No answer provided