

Alerts: Real-Time Governance Across Pencil

Content-aware, blocking governance rules with a clear, powerful UI — live across Editor, Agents, Workflows, Sheets & Feeds

What it is

Alerts is Pencil's real-time governance layer. Workspace admins define trigger-action rules — based on prompt content, model selection, agent, workflow context, or asset scores — and choose the response: a guidance alert (informational, user can continue) or a block (a hard stop until the violation is resolved). Enforcement works at both ends of the creative process: before generation, by evaluating the request itself, and after generation, by scoring the output and gating what can happen to it. A block doesn't just pause the action in front of the user — it prevents the affected content from leaving Pencil entirely, disabling export, download, and sharing to any external destination until the content passes.

Key capabilities

A clear, powerful Alerts UI

- Alerts clearly communicate what triggered the rule, why the action is blocked, and how to resolve it.
- Blocking alerts are visually distinct from informational/warning alerts, reinforcing enforcement vs. guidance — a block stops the action entirely, while a warning lets it proceed but flags the result for review.
- Non-dismissible alerts ensure users must fix the issue before proceeding — alerts cannot be bypassed by closing the popup.

Content-aware alert triggers

- Alerts evaluate user input and configuration — prompt content, selected models, agent, and workflow context — not just button clicks.
- Rules apply across entire capabilities (e.g. all image generators), not individual tools, and automatically extend to future agents without additional setup.

Blocking actions by design

- Generation or execution is intercepted before it starts — no credits are consumed on non-compliant requests.
- The UI guides users toward the correct next step, such as switching to an approved model or adjusting the input.
- Blocked content cannot be exported, downloaded, or shared to any external destination until it passes.

Score-based alerts on generated content

- Alerts can also trigger on Scores, evaluating the quality and compliance of content after generation — not just the request before it.
- Generated content enters an “Evaluating” state while scores are computed; admins set score thresholds that decide what happens next.
- Actions range from informational guidance on low-scoring content to hard enforcement — e.g. keeping Export and Share disabled until the content is regenerated to a passing score.

Where it works

Alerts are live across every creation surface in Pencil:

- **Editor** — legal/compliance disclaimers on image and video generation, plus real-time asset-scoring alerts on the canvas. A failing score can disable Export and Share until resolved.
- **Chat & Agents** — rules evaluate prompt content, selected model, and active agent before a message is sent. Alerts can redirect users to the correct specialized agent or block non-compliant requests outright.
- **Workflows** — rules run node-by-node. A violation highlights the offending node and halts all downstream nodes (“circuit breaker”), so one bad step can't silently consume credits on the steps after it.
- **Sheets & Feeds** — every AI action in a feed or sheet cell (image, video, agent, remove-background) is validated per cell. Blocked cells skip generation and show a red inline badge; warning cells still generate but are flagged yellow. Clicking a badge opens the alert details, and a “Review N alerts” panel lists every flagged cell for quick navigation. A single cell can carry multiple alerts when several generation types are involved, and indicators persist across reload and stay correctly placed through row/column changes and undo/redo.

Why it matters

- Alerts is a governance layer, not a notification system. Teams enforce policies before risk is introduced, while creatives stay productive with clear, in-context guidance.
- **Brand & legal safety** — enforce ethical-model policies, usage disclaimers, and content rules automatically — and guarantee that content failing those rules can never be exported out of Pencil to any external source.
 - **Fewer review cycles** — issues are surfaced while the creator is still working, when they're cheapest to fix.
 - **Credit efficiency** — blocking violations before generation (and halting downstream workflow nodes) prevents wasted generations.
 - **One framework, every surface** — a single trigger-action engine applies consistently across the platform, so new surfaces and agents inherit governance rules automatically.

Availability

Surface	Capability	Status
Editor	Legal/compliance alerts on image & video generation	Live
Editor	Real-time asset-scoring alerts with export blocking	Live
Chat & Agents	Prompt-, model- and agent-based alerts & blocks	Live
Workflows	Node-level alerts with downstream circuit breaker	Live
Sheets & Feeds	Alerts on AI actions in cells, incl. multiple alerts per cell	Live
Scribble	Full alert coverage across Scribble experiences	Coming next

Key details

- Alerts are fully configurable and only trigger when defined conditions are met — workspaces without rules see no change.
- Enforcement operates on requests and generated outputs; no changes are made to the AI models themselves.
- Alert rules are reusable across multiple workflows and future use cases — define once, enforce everywhere.

What's next

- As the platform evolves toward Scribble, the Alerts framework is being extended to it next — bringing the same governance rules to Scribble experiences so coverage stays seamless as teams adopt the new experience. Also on the roadmap:
- Expanded alert conditions across additional inputs and actions.
 - Deeper integration between Alerts and Scores for end-to-end governance.
 - Admin-level visibility and reporting on alert triggers and compliance.
 - Email notifications for alert events.

Getting started: Workspace admins can configure rules under Settings → Alerts / AI Governance. Rules apply immediately to all users in the workspace.
