



INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

AI Data Classification: What Is Safe for ChatGPT & Copilot

Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-02-25 · ai data privacy · generative ai policy · data classification · chatgpt security · llm compliance · sensitive data · copilot safety · gdpr · pii protection · ai

Original article: <https://intuitionlabs.ai/articles/ai-data-classification-chatgpt-copilot-gemini-policies>



In this article

1. Executive Summary
 2. Introduction and Background
 3. 1. Legal and Regulatory Context
 4. 2. OpenAI ChatGPT: Usage Policies and Data Handling
 5. 3. Microsoft Copilot: Workplace Assistant Data Rules
 6. 4. Google Gemini: Generative AI Under the Google Umbrella
 7. 5. Anthropic Claude: The Constitution and Content Guardrails
 8. 6. Comparative Analysis and Case Studies
 9. 7. Data Classification for AI in Practice
 10. 8. Implications and Future Directions
 11. Conclusion
-

Executive Summary

The rapid proliferation of advanced generative AI systems – notably OpenAI's ChatGPT, Anthropic's Claude, Microsoft's Copilot series, and Google's Gemini – has raised urgent questions about **data classification** and content policy. In practice, users must carefully consider *what data* and *which requests* these models can safely handle. All major AI providers impose strict usage policies: for example, OpenAI's **Usage Policies** (effective Oct 2025) expressly forbid prompts that facilitate violence, self-harm, hate, illegal activities, or misuse of personal data ([platform.openai.com](https://platform.openai.com/policies)) ([platform.openai.com](https://platform.openai.com/policies)). Similarly, Google's Generative AI **Prohibited Use Policy** bans content related to child sexual abuse, terrorism, non-consensual intimate imagery, extremism, and hate ([policies.google.com](https://policies.google.com/generativeai)) ([policies.google.com](https://policies.google.com/generativeai)). Microsoft's AI **Code of Conduct** likewise prohibits deceiving or harming individuals, fraud, and unauthorized IP infringement ([www.microsoft.com](https://www.microsoft.com/en-us/ai/code-of-conduct)) ([learn.microsoft.com](https://learn.microsoft.com/en-us/ai/code-of-conduct)). Anthropic's constitutional guidelines for Claude emphasize broad ethical constraints (e.g. *never* aiding in bioweapon creation ([www.anthropic.com](https://www.anthropic.com/constitution))) and filter out attempts to reproduce copyrighted or unlawful content ([support.anthropic.com](https://support.anthropic.com/en/articles/12000000-constitutional-guidelines)) ([www.anthropic.com](https://www.anthropic.com/constitution)).

Sensitive personal and corporate data requires a workflow review before submission: check authorization, purpose, service tier, retention, training settings, and enabled controls. OpenAI's **Usage Policies** prohibit compromising others' privacy by handling private or sensitive information without authorization. Public availability or possession of a document does not by itself establish permission to use it. Minimize identifiers and obtain organizational approval for confidential inputs.

This report provides an exhaustive examination of **data classification and permissible content** for each of Claude, ChatGPT, Copilot, and Gemini. We review the historical development of content policies and data-handling commitments, analyze each vendor's official guidelines (and how they are enforced), and compare across multiple dimensions (e.g. personal data, copyrighted/trade-secret data, regulated advice). Finally, we discuss future directions: emerging regulations (e.g. the EU AI Act), evolving business adoption (enterprise AI with stricter data controls), and technical safeguards (improved filtering, in-model data sanitization).

Introduction and Background

Generative AI chatbots and assistants have seen explosive adoption since 2022. OpenAI's ChatGPT (initially based on GPT-3.5) quickly amassed hundreds of millions of users, while competitors emerged offering specialized variants. Anthropic introduced **Claude** in early 2023, positioning it as a "helpful and harmless" chat assistant guided by a formal *constitution* of ethical rules (www.anthropic.com). Microsoft integrated OpenAI models into its ecosystem via **Copilot** (for Office apps, and earlier GitHub Copilot for coding), emphasizing productivity within enterprise workflows. Google entered the fray with **Gemini** (formerly Bard), leveraging its vast data and search integration. By early 2026, these AI "personal assistants" pervade workplaces, education, and consumer lives. With this ubiquity comes critical scrutiny: *What content and data can users safely feed into these models? And by extension, what content will the models refuse or censor?* This is the core of "data classification" for AI – the categorization of input data by sensitivity or policy domains and the delineation of permissible vs. prohibited uses.

Early on, providers recognized that unconstrained LLMs could produce harmful or inappropriate outputs. OpenAI's first GPT launch included a content-filtering classifier, and Anthropic began its research focusing on safe AI. Over time, *formal usage policies* have been articulated. For OpenAI, every ChatGPT (site or API) use is governed by the **Usage Policies** ([platform.openai.com](https://platform.openai.com/policies)) ([platform.openai.com](https://platform.openai.com/policies)). These universal rules prohibit content that threatens or harasses individuals, incites violence or extremism, depicts sexual violence, or involves illegal activities (drugs, weapons, hacking) ([platform.openai.com](https://platform.openai.com/policies)). They also enshrine privacy concerns: disallowing facial recognition databases or profiling individuals without consent ([platform.openai.com](https://platform.openai.com/policies)). The policies explicitly target manipulative practices – forbidding scams, fraud, or academic dishonesty ([platform.openai.com](https://platform.openai.com/policies)) – and set strict safeguards for minors (user under 18) ([platform.openai.com](https://platform.openai.com/policies)). Google's policies similarly forbid dangerous/illegal content, hate speech, and misinformation (policies.google.com) (policies.google.com). Microsoft has codified its stance in an **AI Code of Conduct**: it bans applications that inflict harm or deception, or that serve prohibited content (learn.microsoft.com).

These policies have implications for **data classification**. Organizations typically label data as *public*, *internal*, *confidential*, or *highly confidential*. Before submitting a document, check permission to disclose it, personal information, licensing, the intended task, and the selected service's data controls. Treat public and user-provided material as candidates for review, rather than automatically safe inputs. Microsoft advises users to own the content or have rights to it ([Microsoft guidance](#)). Use an approved organizational workflow for sensitive information; minimize or redact data where appropriate.

Figure 1 provides an overview of common content/data categories and the broad interplay with major AI policies:

Category	Allowed/Prohibited (Excerpts from Policies)
Violence/Terrorism	Prohibited: No instructions for bomb-making, violent crimes, terrorism (OpenAI, Google) (platform.openai.com) (policies.google.com).
Hate Speech/Extremism	Prohibited: Content that harasses protected groups or promotes extremist ideology is disallowed (OpenAI, Google) (platform.openai.com) (policies.google.com).

Category	Allowed/Prohibited (Excerpts from Policies)
Sexual Content (Adult)	Restrictions depend on the provider, purpose, and context. Consult each provider's policy rather than treating every discussion of sexual content as prohibited.
Sexual Content (Minors)	Prohibited: Any sexual content involving minors of any kind (platform.openai.com) (policies.google.com).
Self-harm/Suicide	Prohibited: Models must not facilitate or encourage self-harm. (platform.openai.com) (policies.google.com) (But <i>supportive</i> responses may be allowed).
Medical/Legal Advice	Provider-specific rules apply. OpenAI prohibits tailored advice requiring a license without appropriate licensed-professional involvement (OpenAI). Google restricts unsupervised automated decisions with material detrimental effects in high-risk domains (Google).
Illicit Substances/Criminal Acts	Prohibited: Instructions for drug synthesis, hacking, or other crime (platform.openai.com) (policies.google.com).
Privacy/Personal Data	Prohibited: Using models to identify or profile someone through PII is banned (platform.openai.com) (policies.google.com). No location tracking.
Fraud/Deception/Impersonation	Prohibited: Generating scams, fake endorsements, or impersonating others without disclosure (platform.openai.com) (policies.google.com).
Copyrighted Material	Check rights and licensing. A supported analysis or transformation workflow does not authorize infringement or unrestricted reproduction (Microsoft guidance).
Harassment/Personal Attack	Prohibited: Insults, bullying, or defamation of individuals/groups (platform.openai.com) (policies.google.com).
Extremist Content Generation	Prohibited: Any content favoring violent extremist acts or ideology.

This table is illustrative; it is a policy overview, not a test of model refusal rates. OpenAI's [Model Spec](#) documents limited transformation of directly supplied restricted or sensitive content. The assistant must not add new disallowed content or refine dangerous details, and prohibited content remains prohibited during transformation.

We examine each platform's nuances in the sections that follow. Critically, beyond policy language, we examine how content filtering is implemented and enforced: engineers testing the models find certain prompts forcibly blocked, while others slip through or are handled via "safe completions." We also consider the privacy implications of *data input*: what information is stored, who can see it, and under what circumstances (for instance, has OpenAI changed how it uses user chats for training ([openai.com](https://openai.com/policies))?). These practical realities show how "classification" of data (by sensitivity or risk) interacts with each model's design.

1. Legal and Regulatory Context

Before using an AI service, identify the jurisdiction, the organization's role, the data involved, the processing purpose, and the applicable contractual safeguards. Escalate unresolved privacy and licensing questions to the organization's responsible reviewers.

Other regulations impose further data classification. The **COPPA** law in the US prohibits collection of data from children under 13 without parental consent. Although ChatGPT's policies bar children from any sexual content (platform.openai.com), responsible use requires not eliciting identifiable info from minors. In practice, OpenAI's initial approach has been to block queries that appear to involve under-18 sexual content (platform.openai.com), and provide parental controls on voice chat. However, organizations are wise to avoid having employees input any personal data of minors.

Health and other regulated data require a use-case review. **HHS** states that HIPAA applies to covered entities and business associates; an individual discussing their own health is not automatically in either role. For organizational PHI processing, verify the exact eligible product and functionality, the executed BAA where required, configuration, and remaining compliance obligations. **OpenAI** explains that a BAA and enabled support do not by themselves make an application HIPAA compliant. **Google's Workspace privacy hub** documents support for HIPAA workloads; check its included functionality and BAA scope rather than assuming coverage for the consumer Gemini app or every feature. Financial-data processing needs its own legal and organizational review. Do not disclose another person's information without authorization.

Intellectual property law also classifies data. Proprietary source code or unpublished technical plans are valuable secrets. Sharing them with an AI might infringe trade-secret protections unless covered by NDAs. Microsoft explicitly warns Copilot users to **"Use licensed or original content"** to prevent IP leakage (www.microsoft.com). Similarly, Anthropic's systems detect and block instructions to replicate existing content (support.anthropic.com), implicitly respecting copyright and trade-secret boundaries.

Lastly, **data classification frameworks in organizations** (public, confidential, secret) should guide AI use. Internal policy should specify which classifications an approved service may process and who can authorize exceptions. OpenAI does not train on organization data by default (**OpenAI business data**). That commitment is separate from retention: ChatGPT workspace retention and API zero data retention are different controls. API zero data retention requires approval and remains subject to endpoint and feature limitations (**API data controls**). Do not infer zero retention from an enterprise subscription alone.

2. OpenAI ChatGPT: Usage Policies and Data Handling

OpenAI's **Usage Policies** govern use of ChatGPT and the API. For data classification, distinguish prohibited privacy abuse from an authorized organizational processing workflow. The policy prohibits attempts to aggregate, monitor, profile, or distribute individuals' private or sensitive information without authorization. These are usage requirements; they do not guarantee a particular response to every prompt.

Content Categories:

From the Usage Policy (platform.openai.com) (platform.openai.com), we can summarize ChatGPT's content classification roughly as:

- **Disallowed:**
 - Threats, harassment, defamation.
 - Self-harm facilitation or encouragement.

- Sexual violence or non-consensual sexual content.
- Extremist/terrorist content or advice.
- Weapons or illicit goods procurement.
- Malicious hacking or IP infringement.
- Gambling with real money.
- Tailored advice requiring a license without appropriate licensed-professional involvement.
- Fraud and deception.
- Manipulating elections or government processes.
- High-risk automated decisions without human oversight (in sectors like healthcare, law).
- **Between Allowed and Caution:**
 - Some content might be allowed but with “safe completions.” For example, a user expressing suicidal thoughts may yield the model giving a crisis hotline on “self-harm” rather than elaborating.
 - **Professional Advice:** General information differs from tailored advice requiring a license. The latter requires appropriate involvement by a licensed professional ([OpenAI policy](#)); professional oversight is not a guarantee of accuracy.
- **Allowed (generally safe):**
 - Non-sensitive factual questions.
 - Creative writing (as long as it’s not disallowed content).
 - Limited transformation or analysis of directly supplied restricted or sensitive content, subject to the Model Spec’s exceptions and limits.
 - Assess educational or historical tasks by their purpose and the applicable policy, rather than treating the topic alone as permission.

The [OpenAI Model Spec](#) explicitly documents a transformation provision. It covers limited analysis or transformation of restricted or sensitive material supplied directly by the user, without adding new disallowed content or refining dangerous details. It does not override other prohibitions, including the prohibition on producing sexual content involving minors. This is a description of intended behavior, not a universal refusal guarantee.

Privacy and Personal Data:

Beyond the content catalog, distinguish four data controls:

- **Training:** Consumer users can turn off “Improve the model for everyone” so new conversations are not used for training ([ChatGPT data controls](#)). OpenAI does not train on business data by default ([business data](#)).

- **Saved chats and workspace retention:** Regular and archived chats remain saved until deletion or removal under an applicable workspace retention policy. Deleted chats are scheduled for permanent deletion within 30 days, with exceptions for de-identification and security or legal obligations ([chat retention](#)).
- **Temporary Chat:** OpenAI may keep a copy for up to 30 days for safety purposes; this is separate from saved-chat retention ([retention documentation](#)).
- **API abuse monitoring:** Logs are retained for up to 30 days by default, with stated exceptions. Eligible customers can seek approved zero data retention or modified abuse monitoring; feature and endpoint limitations can still permit application-state storage ([API data controls](#)). An enterprise ChatGPT subscription does not automatically supply this API control.

3. Microsoft Copilot: Workplace Assistant Data Rules

Microsoft 365 Copilot is a workplace assistant governed by Microsoft 365 controls and contractual commitments. Keep it distinct from GitHub Copilot for coding and consumer Copilot; confirm the exact product and signed-in account before choosing an input workflow. Microsoft's employee guidance advises using content that you own or have rights to ([Microsoft guidance](#)).

Use the applicable product terms and organizational controls to assess a request. Do not assume every prohibited prompt will be refused, or use reframing to evade a restriction.

For workplace use, review the guidance alongside the tenant's enabled controls. The employee blog is guidance with illustrative prompting examples, rather than an incident report. A model refusal or successful completion does not by itself establish permission to process the input.

Data Privacy and Licensing:

Microsoft 365 Copilot processes and stores interaction data under Microsoft 365 contractual commitments; a tenant is not an on-premises processing boundary ([Microsoft privacy documentation](#)). Copilot surfaces organizational data that the user has permission to view, making access management important. Web search can send generated queries to Bing. For connected agents, review each agent's privacy statement, terms, and permissions. Verify these settings before approving confidential workflows.

Purview enforcement depends on configured DLP rules. At the Microsoft 365 Copilot and Copilot Chat policy location, sensitivity-label rules can exclude supported files and emails from processing; sensitive-information-type rules address text entered into prompts. A label alone is not a universal paste blocker: copied text needs an applicable detection rule. Microsoft documents support in Copilot, Copilot Chat, and Word, Excel, and PowerPoint, with rollout and feature limitations. Directly uploaded files are not scanned by this prompt DLP feature. Test the actual tenant configuration and workload before marking a classification as blocked ([Purview documentation](#)).

Consult the applicable product terms rather than assuming that every Microsoft AI service uses identical controls. Protected-material detection is not available in every Copilot scenario ([Microsoft documentation](#)); filters do not certify licensing or confidentiality.

Intellectual Property (Code) Focus:

GitHub Copilot uses coding context, including retrieved codebase context, to generate suggestions. Authorized proprietary-code workflows should be assessed against company approval and plan-specific settings rather than treated as prohibited by an IP filter. GitHub states that it does not use Copilot Business or Enterprise data to train its models. Individual-plan interaction data may be used for training; users can opt out in their account settings. Review code access, exclusions, generated-code licensing, and secret-handling rules before enabling a workflow ([GitHub documentation](#)).

Employee Guidance:

Microsoft's internal blog provides real-world examples. If a user tries an illegal or harmful prompt, Copilot might simply refuse. If the prompt contains a blocked keyword, Copilot suggests rephrasing. For instance, "Write a motivational speech by Satya Nadella" is blocked (it falls under impersonation), but Copilot will suggest "Write a motivational talk **inspired by** the leadership style of Satya Nadella" (www.microsoft.com). Similarly, if you request Disney quotes, Copilot says no, but it will offer analogous metaphors (www.microsoft.com). These are illustrative examples, not a universal legal test. Reframing must preserve a permitted purpose and respect the applicable rules. The key takeaway is that Copilot's restrictions are meant to **preserve corporate brand and security**, not to throttle creativity. When it "says no," it usually provides an explanation or alternative, keeping the user on track.

In summary, use only information you are authorized to disclose in an approved workflow. Verify permissions, configured DLP rules, retention, web search, and agent access before submitting sensitive content.

4. Google Gemini: Generative AI Under the Google Umbrella

Google's generative AI offerings (initially called Bard, now **Gemini**) follow Google's own AI Principles and policies. Google emphasizes similar themes: *safety, fairness, and privacy*. The publicly documented **Generative AI Prohibited Use Policy** (Dec 2024 update) is a concise enumeration of restrictions (policies.google.com). It divides forbidden uses into categories:

- **Dangerous/Illegal:** content that relates to or facilitates violent extremism, terrorism, self-harm, child abuse, or other illegal activities is disallowed (policies.google.com).
- **Security:** content that facilitates hacking, phishing, malware, or censorship circumvention is prohibited (policies.google.com).
- **Harassment/Hate:** promotion of hate speech, harassment, or violence is barred (policies.google.com).
- **Sexual:** explicit sexual content (for pornographic intentions) is disallowed (policies.google.com).
- **Misinformation/Deception:** fraud, impersonation, and misleading claims are not allowed (policies.google.com). Notably, Google also specifies you cannot misrepresent AI content as human-written without permission (policies.google.com), which touches on provenance.
- **Privacy/Infringement:** using AI to violate privacy or IP is forbidden (policies.google.com) (policies.google.com). For example, generating non-consensual intimate images or using someone's likeness to deceive is disallowed (policies.google.com) (policies.google.com).

These align closely with OpenAI's and Microsoft's restrictions. Google also explicitly calls out “automated decisions that affect rights” without human supervision in critical domains as prohibited (policies.google.com), anticipating the EU AI Act's high-risk provisions.

Privacy and Data Usage:

Distinguish Gemini under a Google Workspace agreement from the consumer Gemini app. Google's [Workspace privacy hub](#) states that content is not human reviewed or used for generative AI model training outside the domain without permission. This is a scoped Workspace commitment; check the account, service, enabled features, and retention configuration before submitting internal documents.

For consumer use, Google's [Gemini Apps Privacy Hub](#) cautions against entering confidential information that users would not want reviewers to see or Google to use to improve its services. Review activity and privacy settings rather than extending Workspace protections to every Google account or Gemini interaction.

Gemini's Allowed Content and Capabilities:

Gemini shares ChatGPT's broad capabilities (text chat, image understanding, etc.), so most benign or factual tasks are allowed. It can summarize text, answer general knowledge questions, write creative content, and even generate images (Gemini Pro). Providers stress **contextual sensitivity**. For instance, Google's support page on the policy emphasizes that it *expects user responsibility*. It warns that automated systems and human reviewers will detect misuse of Gemini (support.google.com). If a user attempts to bypass the restrictions (such as via prompt injection), Google may flag and block the content. Repeated violations could even lead to suspending the Google account (support.google.com).

Comparison and Enforcement:

Written policies describe acceptable use; they do not establish comparative refusal rates. Evaluate the intended task under the relevant Google policy, and test the specific product and settings before relying on its controls.

For organizational use, approve a specific Gemini product and workflow after reviewing data access, training permissions, retention, and any regulated-data requirements. A model's willingness to answer is not a confidentiality or compliance certification.

5. Anthropic Claude: The Constitution and Content Guardrails

Anthropic's Claude is notable for its **Constitutional AI** approach and emphasis on ethics. Unlike the others, Anthropic publicly released *Claude's Constitution* – a lengthy document of hierarchical values (www.anthropic.com) (www.anthropic.com). We will not reproduce it in full, but highlight points most relevant to input data:

- **Safe Helpful Assistant:** Claude is designed to be “exceptionally helpful, while also being honest, thoughtful, and caring” (www.anthropic.com). It is instructed to be cautious about providing disallowed advice.

- **Hard Safety Constraints:** The constitution lists certain absolute no-goes. For example, Claude should “never provide significant uplift to a bioweapons attack” (www.anthropic.com). It should also refuse to knowingly assist someone who intends harm. These constraints mirror the “illegal content” bans of other firms, and show Anthropic’s explicit focus on **not amplifying harmful use-cases**.
- **Consent and Privacy:** Read Anthropic’s published Usage Policy alongside the constitution. Its enumerated usage standards and high-risk requirements, rather than community impressions, should guide deployment review ([Anthropic policy](#)).

Usage policies describe permitted use; they do not provide a measured ranking of Claude’s refusal rates against ChatGPT. Evaluate a specific task against Anthropic’s policy and test the relevant model and product when behavior matters.

Anthropic’s *official* stance on sensitive personal data is conveyed via its policies and FAQs rather than the constitution text. The Claude FAQ explicitly warns: “**We recommend being thoughtful about sharing highly sensitive personal details**” such as Social Security numbers, medical records, passwords, or confidential business documents (support.anthropic.com). While Anthropic uses privacy techniques (obfuscation, auditing) when reviewing user chats for training (support.anthropic.com), it still counsels users to avoid pasting in raw sensitive info. Importantly, Claude can analyze text containing personal data (e.g. “Summarize this letter from my bank”), but the user should remove or anonymize names, SSNs, etc., before inputting.

Data Use and Retention:

For commercial products such as Claude for Work and the Anthropic API, Anthropic does not use inputs or outputs to train its models by default. Explicit feedback or another choice to share data can change that ([Anthropic privacy documentation](#)).

Consumer Claude users have a model-training choice, separate from deleting saved conversations. Anthropic’s [consumer retention documentation](#), dated July 1, 2026, distinguishes these controls:

- **Saved chats and deletion:** Turning off model improvement does not automatically delete saved conversations. When a user deletes a conversation, it leaves chat history immediately and is deleted from back-end storage within 30 days, subject to the exceptions below.
- **Training pipelines:** With model improvement enabled, de-identified data may be retained for up to five years in training pipelines, covering new or resumed chats and coding sessions after enabling the setting.
- **Exceptions:** Policy-flagged inputs and outputs may be retained for up to two years, safety classification scores for up to seven years, and feedback-related data for five years. Legal requirements, disputes and policy enforcement can justify further retention. Safety review and explicit feedback can also permit training despite opting out ([Privacy Policy](#)).

Anthropic’s [August 2025 announcement](#) described a 30-day period for users declining training. Treat that as historical context; use the current deletion-based rule and exceptions. Consumer settings do not replace commercial product terms.

Usage Examples and Guidelines:

Anthropic publishes explicit usage standards and requirements for consumer-facing high-risk use cases ([Usage Policy](#)). For covered advice, recommendations, or subjective decisions affecting individuals or consumers, its policy requires review by a qualified professional before dissemination or finalization. Do not label all medical or legal research, drafting, or general information as uniformly prohibited; assess the use case and applicable safeguards. The constitution and written policy do not establish that Claude is more restrictive than another provider.

6. Comparative Analysis and Case Studies

To synthesize the above, we compare **what each model permits versus forbids**, and examine illustrative examples.

Content Categories Comparison: The table below summarizes key categories of data/content and the stance of each AI system (a policy overview; not a guarantee of observed behavior):

Content / Data Type	ChatGPT (OpenAI)	Copilot (Microsoft)	Gemini (Google)	Claude (Anthropic)
Hate / Harassment	Disallowed (platform.openai.com) (policies.google.com)	Disallowed (www.microsoft.com) (policies.google.com)	Disallowed (policies.google.com)	Disallowed (ethical guidelines)
Violence / Extremism	Disallowed (platform.openai.com) (policies.google.com)	Disallowed (www.microsoft.com) (policies.google.com)	Disallowed (policies.google.com)	Disallowed (www.anthropic.com)
Self-harm / Suicide	Disallowed (no facilitation) (platform.openai.com)	Disallowed	Disallowed	Typically disallowed/help outweighs harm
Sexual Content (Adults)	Review the applicable content rules and context.	Review the applicable content rules and context.	Review the applicable content rules and context.	Review the applicable content rules and context.
Sexual Content (Minors)	Disallowed (platform.openai.com)	Disallowed	Disallowed (policies.google.com)	Disallowed
Illegal Instructions (drugs/weapons)	Disallowed (platform.openai.com) (www.anthropic.com)	Disallowed (www.microsoft.com)	Disallowed (policies.google.com)	Disallowed (www.anthropic.com)
Privacy / PII Exploitation	Disallowed (no profiling/shedding PII) (platform.openai.com)	Disallowed (no impersonation) (www.microsoft.com)	Disallowed (no facial recognition/biometrics) (policies.google.com)	Disallowed

Content / Data Type	ChatGPT (OpenAI)	Copilot (Microsoft)	Gemini (Google)	Claude (Anthropic)
Defamation / Misinfo	Disallowed (no libel/gross disinfo) (platform.openai.com)	Disallowed (no fraud/impersonation) (www.microsoft.com)	Disallowed (no fraud/impersonation) (policies.google.com)	Disallowed
Financial Advice	General information differs from tailored advice requiring a license; the latter requires appropriate licensed-professional involvement (policy).	Review the specific product terms and oversight requirements.	Do not make unsupervised automated decisions with material detrimental effects in high-risk domains (policy).	Apply consumer-facing high-risk requirements, including qualified-professional review where covered (policy).
Medical Advice	General information differs from tailored advice requiring a license; the latter requires appropriate licensed-professional involvement (policy).	Review the specific product terms and oversight requirements.	Do not make unsupervised automated decisions with material detrimental effects in high-risk domains (policy).	Apply consumer-facing high-risk requirements, including qualified-professional review where covered (policy).
Copyrighted Content (generation)	Do not infringe others' rights; a filter is not a legal clearance.	Check rights; protected-material detection is scenario-dependent.	Check rights and applicable policy.	Check rights and applicable policy.
Copyrighted Content (analysis)	Limited transformation of directly supplied material has policy limits.	Check authorization and licensing.	Check authorization and licensing.	Check authorization and licensing.
Private Business Data (trade secrets)	Use only an approved workflow; verify training and retention settings.	Verify tenant permissions, DLP, web search, and agents.	Distinguish approved Workspace services from consumer use.	Distinguish commercial controls from consumer settings.
Public Data (news, wiki)	Check privacy, rights, and purpose.	Check privacy, rights, and purpose.	Check privacy, rights, and purpose.	Check privacy, rights, and purpose.

Table 1. Summary of permitted vs. disallowed content by model (Illustrative).

Interpretation: Distinguish prohibited uses from model behavior. Written policies are not a guarantee that every prohibited prompt will be refused or every permissible task will be answered. General professional-topic information differs from tailored advice and high-risk automated decisions; apply the provider-specific requirements above. Content filters do not certify legality, licensing, or confidentiality.

The main differences arise around **intellectual property** and **private corporate data**. A user-provided document is not automatically licensed for every requested use. OpenAI's documented transformation provision has limits ([Model Spec](#)), and Microsoft notes that protected-material detection is not available in every Copilot scenario ([Microsoft documentation](#)). For confidential inputs, verify the specific product, approved workflow, data-access controls, training settings, and retention policy.

Guidance and a Hypothetical Workflow

Microsoft's employee blog provides illustrative prompting guidance, rather than an incident report. Its practical recommendation is to review AI-generated content before publication ([Microsoft guidance](#)).

Hypothetical example: A bank considering AI-assisted research on customer records should first determine the data classification, processing purpose, authorization, and applicable legal requirements. Its reviewers should approve the specific service, account, contractual scope, retention, and access controls. Masking names alone should not be treated as proof that a dataset is anonymous or approved for disclosure. This is an illustration, not a report about a named or unnamed bank's deployment.

7. Data Classification for AI in Practice

In practical terms, data classification for AI involves mapping traditional data categories (public, sensitive, private, confidential) to AI-specific risks and usage rules. Here we propose a **Data Sensitivity vs. Model Usage** matrix to guide safe input:

Data Type / Classification	ChatGPT consumer	ChatGPT Enterprise	Microsoft 365 Copilot	Gemini	Claude
Public / licensed content	Check privacy, rights, and purpose.	Check privacy, rights, and purpose.	Check privacy, rights, and purpose.	Check privacy, rights, and purpose.	Check privacy, rights, and purpose.
Personal information	Minimize identifiers; verify authorization and settings.	Approve the use case and workspace controls.	Approve permissions and applicable DLP rules.	Distinguish consumer and Workspace controls.	Distinguish consumer and commercial controls.
Confidential / trade-secret information	Do not submit through an unapproved workflow.	Verify approval, training terms, retention, and access.	Verify tenant permissions, DLP, web search, and agents.	Verify the approved service, agreement, and settings.	Verify the approved service, agreement, and settings.
Organizational PHI	Do not assume eligibility or BAA coverage.	Verify eligible functionality, executed BAA where required, and configuration.	Verify eligible functionality, contract scope, and organizational approval.	Verify Workspace HIPAA included functionality and BAA scope; do not infer consumer-app coverage.	Verify eligible functionality and contract scope before processing.
Financial personal data	Review applicable law, purpose, authorization, and controls.	Review applicable law, purpose, authorization, and controls.	Review applicable law, purpose, authorization, and controls.	Review applicable law, purpose, authorization, and controls.	Review applicable law, purpose, authorization, and controls.
Copyrighted material	Verify rights and the requested use.	Verify rights and the requested use.	Verify rights and the requested use.	Verify rights and the requested use.	Verify rights and the requested use.

Data Type / Classification	ChatGPT consumer	ChatGPT Enterprise	Microsoft 365 Copilot	Gemini	Claude
Malicious / prohibited uses	Follow the applicable usage policy.	Follow the applicable usage policy.	Follow the applicable usage policy.	Follow the applicable usage policy.	Follow the applicable usage policy.

Table 2. Data sensitivity vs. AI model input guidelines (illustrative).

Key points from Table 2:

- **Public and user-provided content:** Check privacy, authorization, licensing, and the intended task. Public availability does not certify safe input.
- **Personal and confidential information:** Minimize what you submit and obtain approval for the specific service and account. Model acceptance does not establish lawful disclosure or organizational permission.
- **Workplace Copilot:** Review Microsoft 365 cloud contractual commitments, user permissions, web queries, and connected agents; avoid assuming processing occurs behind a corporate firewall ([Microsoft documentation](#)).
- **Regulated health data:** HIPAA applies to covered entities and business associates ([HHS](#)). For organizational PHI, verify the eligible product/functionality, executed BAA where required, configuration, and remaining obligations. A BAA alone does not establish compliance ([OpenAI](#)). Google documents Gemini support for specified Workspace HIPAA workloads; consult the included-functionality scope ([Google](#)). Personal use is not automatically organizational PHI processing, and financial-data rules need a separate review.
- **Content licensing and prohibited uses:** Check the requested use, not only the source of the text. OpenAI's transformation provision is limited and does not override other prohibitions ([Model Spec](#)). Filters do not certify rights or guarantee a refusal.

Use this matrix as a review workflow rather than a legal or product-wide permission chart. Verify the tenant or account's actual controls before treating a classification as allowed or blocked.

8. Implications and Future Directions

The emerging field of **AI data classification** is dynamic. Provider policies, product capabilities, and legal obligations change over time. But it's crucial to anticipate how things evolve:

- **Regulatory Changes:** The [European Commission's AI Act timeline](#) records entry into force on 1 August 2024, application of prohibited-practice and AI-literacy provisions from 2 February 2025, and governance and general-purpose-AI obligations from 2 August 2025. The Act became applicable on 2 August 2026 with exceptions and transition arrangements. Check the current Commission timeline and the specific provision that covers a deployment; do not infer a blanket ban on every biometric use.
- **Corporate Policies:** Define approved products, permitted data classifications, authorization, retention, and access controls. For regulated workloads, review legal obligations and contract scope before approving a workflow; do not treat an enterprise plan or BAA as blanket permission.

- **Technical Safeguards:** On the tech side, ongoing research aims to improve content filters. For example, OpenAI's Moderation API is free for developers to check outputs. Providers train their models to self-filter more accurately. One research trend is *privacy-preserving prompt sanitization* – automatically detecting and redacting sensitive info before it's sent to the model (support.anthropic.com). Several startups already offer prompt-scrubbing services for compliance. We should expect integrated features (like an “Anonymize” button) soon.
- **Contributor Credits and Content Provenance:** The U.S. [Blueprint for an AI Bill of Rights](#) is non-binding and does not constitute U.S. government policy. It does not establish a watermarking or origin-metadata mandate. For any asserted provenance obligation, identify the applicable legal provision and its scope; distinguish voluntary transparency practices from legal requirements.
- **Comparison of Models:** Assess provider rules separately from measured model behavior. For deployment review, record the model version, product, configuration, prompts, and evaluation method; do not infer comparative refusal rates from policy text or community anecdotes.
- **Ethical Debates and Public Perception:** There is ongoing debate about how stringent policies should be. Some new users complain that content filters are “too restrictive,” flagging benign content erroneously (community.openai.com). Others argue they are insufficient, pointing to harmful outputs that still slip through. Regulators and ethicists will continue to shape norms. For instance, if an AI system inadvertently discloses protected class data (like race or health) from a seemingly innocuous query, that will spark updates. The mere act of classifying data (e.g. labeling something “sensitive”) can influence model behavior; research into conditional training or prompt weighting based on data class is emerging.
- **Global Variations:** Identify the jurisdictions governing the user, organization, affected people, and processing activity. Obtain a use-case-specific review rather than assuming one classification scheme establishes permission everywhere.
- **User Responsibilities and Education:** Finally, there is a human factor. Users must be educated on data classification in AI contexts. Just as firms teach employees not to click suspicious emails, training will include “don't paste customer names into ChatGPT.” Some organizations are developing internal AI governance frameworks. Experts suggest a “second pair of eyes” approach: outputs or sensitive inputs should be reviewed by humans, especially in high stakes scenarios. Indeed, one key principle in all policies is “*Keep humans in the loop.*” The Microsoft blog emphasizes: “**Don't publish without review**” (www.microsoft.com). This echoes a general data classification rule: **even if the AI allows it, consider the risk.** The models can handle sensitive words, but just because “we can input this data” doesn't mean “we should.” Users must classify their own data and apply judgment.

Conclusion

The landscape of **data classification for generative AI** is evolving rapidly. Across ChatGPT, Copilot, Gemini, and Claude, the practical task is to evaluate the intended use and the input's sensitivity alongside the applicable service terms. Review authorization, purpose, licensing, service tier, retention, training settings, and enabled controls. Public facts, user-generated text, and anonymized samples still require that review; neither public availability nor model acceptance establishes permission.

Product-specific controls matter. OpenAI does not train on business data by default ([OpenAI business data](#)); that is separate from retention and legal compliance. Distinguish consumer settings from organizational agreements, and verify the controls in the actual account or tenant. **“What you can put”** depends on the approved workflow and applicable obligations, as well as content rules.

Looking ahead, data classification for AI will only become more formalized. Monitor applicable legal provisions and their staged timelines when reviewing a deployment. Industry best practices are emerging: strong authentication, prompt filtering, anonymization tools, and user training. Technologically, models may evolve to automatically tag inputs by privacy level or risk. Ethically, models themselves may gain awareness of the sensitivity of data (e.g. saying “I’m not allowed to discuss this personal matter” on their own).

The practical principle is to submit only data you are authorized to use in an approved workflow. Minimize sensitive information, review outputs, and check the current provider terms and organizational requirements before processing confidential or regulated material. Classification supports a documented decision; it does not guarantee safety, legality, or model behavior.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. <https://platform.openai.com/docs/usage-guidelines>
2. <https://policies.google.com/terms/generative-ai/use-policy>
3. <https://www.microsoft.com/insidetrack/blog/helping-our-employees-when-microsoft-365-copilot-says-no-you-cant-do-that/>
4. <https://learn.microsoft.com/en-us/legal/ai-code-of-conduct?context=%2Fazure%2Fai-services%2Fagents%2Fcontext%2Fcontext>
5. <https://www.anthropic.com/constitution>
6. <https://support.anthropic.com/en/articles/10023638-why-am-i-receiving-an-output-blocked-by-content-filtering-policy-error>
7. <https://openai.com/policies/usage-policies/>
8. <https://model-spec.openai.com/2025-12-18.html>
9. <https://openai.com/business-data>
10. <https://www.hhs.gov/hipaa/for-professionals/covered-entities/index.html>
11. <https://help.openai.com/en/articles/8660679-getting-a-business-associate-agreement-for-the-openai-api>
12. <https://support.google.com/a/answer/15706919?hl=en>
13. <https://openai.com/business-data/>
14. <https://developers.openai.com/api/docs/guides/your-data>
15. <https://help.openai.com/en/articles/7730893-data-controls-in-chatgpt>
16. <https://help.openai.com/en/articles/8983778-chat-and-file-retention-in-chatgpt>
17. <https://learn.microsoft.com/en-us/microsoft-365/copilot/microsoft-365-copilot-privacy>
18. <https://learn.microsoft.com/en-us/purview/dlp-microsoft365-copilot-location-learn-about>
19. <https://github.com/features/copilot>
20. <https://support.google.com/gemini/answer/13594961?hl=en>

21. <https://support.google.com/gemini/answer/16625148?hl=en>
22. <https://www.anthropic.com/legal/archive/22742366-2ef0-4c7a-a833-6523f10d3944>
23. <https://support.anthropic.com/en/articles/8325621-i-would-like-to-input-sensitive-data-into-my-chats-with-claude-who-can-view-my-conversations>
24. <https://privacy.claude.com/en/articles/7996868-is-my-data-used-for-model-training>
25. <https://privacy.claude.com/en/articles/10023548-how-long-do-you-store-my-data>
26. <https://www.anthropic.com/legal/privacy>
27. <https://www.anthropic.com/news/updates-to-our-consumer-terms>
28. <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>
29. <https://bidenwhitehouse.archives.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf>
30. <https://community.openai.com/t/content-policies-are-downright-crippling/1243658>

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore [AI enablement](#), [pharma and biotech software development](#), [data engineering and BI](#), and [Veeva services](#). [Contact IntuitionLabs](#) to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: <https://intuitionlabs.ai>

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.