AI Data Classification: What Is Safe for ChatGPT & Copilot

By Adrien Laurent, CEO at IntuitionLabs • 2/25/2026 • 45 min read

ai data privacy generative ai policy data classification chatgpt security llm compliance sensitive data copilot safety gdpr pii protection

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 ([1] platform.openai.com) ([2] platform.openai.com). Similarly, Google's Generative AI **Prohibited Use Policy** bans content related to child sexual abuse, terrorism, non-consensual intimate imagery, extremism, and hate ([3] policies.google.com) ([4] policies.google.com). Microsoft's AI **Code of Conduct** likewise prohibits deceiving or harming individuals, fraud, and unauthorized IP infringement ([5] www.microsoft.com) ([6] learn.microsoft.com). Anthropic's constitutional guidelines for Claude emphasize broad ethical constraints (e.g. *never* aiding in bioweapon creation ([7] www.anthropic.com)) and filter out attempts to reproduce copyrighted or unlawful content ([8] support.anthropic.com) ([7] www.anthropic.com).

Across all models, *sensitive personal data* (e.g. identification, health records, financial details) is heavily restricted. OpenAI explicitly disallows profiling individuals or using biometric data (and warned that ChatGPT "rejects requests for private or sensitive information about people" ([9] techcrunch.com)), Microsoft bans impersonation or PII misuse ([5] www.microsoft.com), Google forbids using biometric or personal data without consent ([10] policies.google.com), and Anthropic advises users not to input highly confidential personal or corporate data (financial records, passwords, trade secrets, etc.) even with its privacy protections ([11] support.anthropic.com). Industry analyses illustrate the stakes: for instance, one study found nearly 11% of employee prompts to ChatGPT contained *confidential* information ([12] captaincompliance.com) (unio.digital), and "shadow AI" breaches (unauthorized AI tool use) accounted for roughly 20% of corporate breaches in 2025 ([13] captaincompliance.com).

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). We present case studies (such as the Italian GDPR ruling fining OpenAI €15M ([14] www.lewissilkin.com), leaked ChatGPT conversation analysis ([15] www.safetydetectives.com), and Microsoft's Copilot "we can't do that" blog ([16] www.microsoft.com)) to illustrate real-world implications. Statistical findings (like the fraction of sensitive inputs and data breach figures ([13] captaincompliance.com) ([15] www.safetydetectives.com)) and expert commentary are integrated throughout. 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 (based on GPT-3.5 and GPT-4) 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 ([17] 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** ([1] platform.openai.com) ([2] platform.openai.com). These universal rules prohibit content that threatens or harasses individuals, incites violence or extremism, depicts sexual violence, or involves illegal activities (drugs, weapons, hacking) ([1] platform.openai.com). They also enshrine privacy concerns: disallowing facial recognition databases or profiling individuals without consent ([18] platform.openai.com). The policies explicitly target manipulative practices – forbidding scams, fraud, or academic dishonesty ([19] platform.openai.com) – and set strict safeguards for minors (user under 18) ([20] platform.openai.com). Google's policies similarly forbid dangerous/illegal content, hate speech, and misinformation ([3] policies.google.com) ([4] 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 ([6] learn.microsoft.com). All providers stress that their model **territory** excludes truly personal (private individual) or misused content – for example, ChatGPT may refuse to reveal personal info about private individuals without consent ([9] techcrunch.com).

These policies have implications for **data classification**. Organizations typically label data as *public, internal, confidential,* etc., and training/usage must respect those labels. With AI chatbots, the rule-of-thumb is usually: **public or user-provided material is safe, but confidential/private content is off-limits**. For instance, Anthropic warns against sharing "highly sensitive personal details" like Social Security numbers, health records, and private credentials ([11] support.anthropic.com). Similarly, Microsoft advises Copilot users to input only "licensed or original content" to avoid IP breaches ([21] www.microsoft.com). This roughly corresponds to data classifications: public domain or open-source data is fine; personal data (PII); sensitive data (financial, health); and secret/classified data should *not* be fed into these black-box models.

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

Category	Allowed/Prohibited (Excerpts from Policies)
Violence/Terrorism	***Prohibited***: No instructions for bomb-making, violent crimes, terrorism (OpenAI, Google) ([1] platform.openai.com) ([3] policies.google.com).
Hate Speech/Extremism	***Prohibited***: Content that harasses protected groups or promotes extremist ideology is disallowed (OpenAI, Google) ([1] platform.openai.com) ([4] policies.google.com).
Sexual Content (Adult)	***Prohibited***: Explicit pornographic or exploitative sexual descriptions/action ([22] platform.openai.com) ([4] policies.google.com).
Sexual Content (Minors)	***Prohibited***: Any sexual content involving minors of any kind ([23] platform.openai.com) ([3] policies.google.com).
Self-harm/Suicide	***Prohibited***: Models must not facilitate or encourage self-harm. ([24] platform.openai.com) ([3] policies.google.com) (But *supportive* responses may be allowed).
Medical/Legal Advice (Unlicensed)	***Prohibited***: Tailored health or legal advice without qualified oversight ([25] platform.openai.com) ([26] policies.google.com).
Illicit Substances/Criminal Acts	***Prohibited***: Instructions for drug synthesis, hacking, or other crime ([27] platform.openai.com) ([3] policies.google.com).
Privacy/Personal Data	***Prohibited***: Using models to identify or profile someone through PII is banned ([18] platform.openai.com) ([10] policies.google.com). No location tracking.
Fraud/Deception/Impersonation	***Prohibited***: Generating scams, fake endorsements, or impersonating others without disclosure ([19] platform.openai.com) ([26] policies.google.com).
Copyrighted Material	***Disallowed (for generation)***: Verbatim recreations of copyrighted text are blocked ([8] support.anthropic.com); minor transformation (summary, translation) may be permissible if user-provided.
Harassment/Personal Attack	***Prohibited***: Insults, bullying, or defamation of individuals/groups ([1] platform.openai.com) ([4] policies.google.com).
Extremist Content Generation	***Prohibited***: Any content favoring violent extremist acts or ideology.

This table is illustrative (policies use broader language), but it highlights that *none of the models allow blatantly illegal or harmful content*. Even transformation of disallowed content is limited: for example, Claude's filter will block attempts to reproduce copyrighted works ([8] support.anthropic.com), and OpenAI's policy labels competing tasks (e.g. "analysis of user-provided text") as safe only when strictly for transformation, not for new generation.

We will delve into 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 ([28] openai.com)?). These practical realities show how "classification" of data (by sensitivity or risk) interacts with each model's design.

1. Legal and Regulatory Context

Generative AI does not exist in a vacuum: it operates under broad legal frameworks that effectively classify data by sensitivity. In particular, **privacy laws** (GDPR in Europe, CCPA/CPRA in California, CPPL in China, etc.) make input of personal data into service "of the cloud" a legal issue. For example, the GDPR defines *personal data* broadly as any information relating to an identified or identifiable person. Under GDPR, processing personal data needs a legal basis. This became a flashpoint in early AI history: in 2023-24 the Italian Data Protection Authority (Garante) investigated ChatGPT. It concluded that OpenAI had been "mass collecting and storing personal data for training" without a proper legal basis ([29] www.lewissilkin.com) ([30] www.lewissilkin.com). Consequently, OpenAI was fined €15 million in Dec 2024 for multiple GDPR violations, and ordered to undertake a public transparency campaign ([14] www.lewissilkin.com) ([30] www.lewissilkin.com). The Garante cited "lack of appropriate legal basis" for data used in AI training and failure to inform users about data collection ([29] www.lewissilkin.com) ([30] www.lewissilkin.com). This enforcement action underscores that, under European law, **sensitive or personal data should not be carelessly fed into AI services without consent or necessity**. (For instance, if a French or Italian user inputs Spanish medical records into ChatGPT, that processing would likely violate GDPR absent safeguards.)

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 ([23] 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 ([23] platform.openai.com), and provide parental controls on voice chat. However, organizations are wise to avoid having employees input any personal data of minors.

Special category data (health, biometrics, genetics, religion, etc.) also commands strict treatment. Under GDPR these require stronger justification. Entering a patient's health record into an AI to get advice would require HIPAA or equivalent protections. ChatGPT Enterprise offers a HIPAA Business Associate Agreement (BAA) for qualified customers ([31] openai.com), acknowledging that medical records as input are often subject to HIPAA. In contrast, ChatGPT free (and presumably Google's public Gemini, Anthropic's public CLAUDE) have no HIPAA mechanism; thus any private health data would likely breach privacy laws. Indeed, all providers caution that they are not substitutes for licensed professionals – implicitly recognizing that giving truly personal medical advice is disallowed ([25] platform.openai.com) (and lack of oversight could even contravene medical regulations).

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 ([21] www.microsoft.com). Similarly, Anthropic's systems detect and block instructions to replicate existing content ([32] support.anthropic.com), implicitly respecting copyright and trade-secret boundaries. Nonetheless, users must be judicious: a leaked example involved Samsung engineers inadvertently pasting secret chip designs into a public AI, triggering internal security alarms.

Lastly, **data classification frameworks in organizations** (public, confidential, secret) should guide AI use. Many enterprises now forbid any *confidential* material in ChatGPT unless absolutely controlled. For example, internal policies may mandate that only "Public" or "Internal" emails can be summarized by an LLM; anything marked "Confidential" is off-limits. This aligns with data retention realities. OpenAI's business documentation makes it clear: *"by default, we do not use organization's data from ChatGPT Enterprise or API for training"* ([28] openai.com). In other words, with a paid enterprise license one can configure *zero retention* and avoid model-training usage, mitigating legal risk. But under a consumer account, all inputs might be stored and reviewed (with de-identification) for service improvement, per OpenAI's privacy policy. Thus the legal/regulatory landscape is pushing companies to classify their data empirically: **share only unsensitive data (public info, anonymized samples)** with public AI, and restrict anything that could violate privacy or IP laws.

2. OpenAI ChatGPT: Usage Policies and Data Handling

OpenAI's ChatGPT (including the web app and API) sets the benchmark for content policies. As of late 2025, OpenAI publishes a **universal Usage Policy** (effective Oct 29, 2025 ([1] platform.openai.com)) applying to all interactions. Key points (as noted above) include bans on violence, hate, sexual exploitation, illegal advice, and privacy violations ([1] platform.openai.com) ([2] platform.openai.com). Importantly for data classification, OpenAI requires users not to input private data of others without consent. The policy explicitly forbids "the evaluation or classification of individuals based on their social behavior, personal traits, or biometric data (social scoring, profiling, sensitive attributes)" ([33] platform.openai.com). This means tasks like "tell me Kerry's blood type" or "analyze this person's behavior from their social media" would be disallowed. In practice, ChatGPT aggressively refuses prompts that might elicit personal data about a private individual, stating it cannot comply without apparent consent. OpenAI emphasizes: *"We want our AI to learn about the world, not about private individuals."* ([9] techcrunch.com).

Content Categories:

From the Usage Policy ([1] platform.openai.com) ([2] 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.
- Unlicensed legal or medical advice.
- Academic cheating (fraud, dishonesty).
- 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**: Pure emotional or general health/law information is permitted (e.g. "What are symptoms of flu?"), but anything requiring customization (e.g. diagnosing a specific person) is barred.
- **Allowed (generally safe)**:
- Non-sensitive factual questions.
- Creative writing (as long as it's not disallowed content).
- Language translation (even of user-provided disallowed content under transformation rules) – ChatGPT allows summarizing or rewriting user-supplied text even if it contains some disallowed material, since it is considered a user's *transformation request*.
- Educational or historical discussions of disallowed topics (e.g. a factual account of war) are typically allowed under freedom-of-speech exceptions.

This is reflected in the model's behavior. In its point of view, it will *refuse* requests that *create* disallowed content; but if you supply disallowed content yourself and ask for analysis or translation, it will often do so (this is known as the "transformation exception" in AI policy discussions, though not formally stated in OpenAI's docs). For example, one can copy-paste a copyrighted image and ask ChatGPT to describe it; it will often do so, since you provided the data. But if you ask it to generate the Disney logo from scratch, it will refuse.

Privacy and Personal Data:

Beyond the content catalog, OpenAI's handling of **user data** is crucial. By default, ChatGPT's free and Plus versions **do** use input and output to train models, unless the user explicitly opts out in settings ([28] openai.com). (GPT-4o, GPT-4.1 models ahead had occasional data-handling differences, but the default is training contributions.) This means anything you enter *could* become training fodder for future models – with anonymization steps. In July 2025, OpenAI clarified it does *not* train on ChatGPT Enterprise, Team, or EDU data unless *explicitly opted-in* ([28] openai.com). However, even Enterprise chat is retained in logs by default for 30 days for abuse monitoring ([34] captaincompliance.com). Only select customers can set retention to zero and otherwise prevent training. Thus, from a data-classification perspective:

- **OpenAI Free ChatGPT** = "high risk": do *not* input confidential or personally identifiable data here (since it will be stored and potentially used in training).
- **ChatGPT Enterprise/Team** = "safer": by agreement, data isn't used for training absent opt-in ([28] openai.com), and can be confined to chosen regions ([35] openai.com), but one must still exercise caution and comply with corporate data policies.

Case: Leaked Conversations

The hazards of improper inputs can be illustrated by real incidents. In August 2025, OpenAI enabled a feature allowing users to share chats publicly. As a result, nearly 1,000 ChatGPT conversations were found publicly indexed, revealing users' private inputs ([36] www.safetydetectives.com). A SafetyDetectives analysis of these leaks (over 43 million words) found a staggering amount of **sensitive content**: people had divulged full names, addresses, financial info, medical conditions, family dynamics, even suicidal thoughts and hate speech ([15] www.safetydetectives.com) ([37] www.safetydetectives.com). Notably, 53.3% of the text came from just ~100 long dialogues, but almost every category of sensitive information appeared. The analysis highlighted that users blindly trust these AIs, often sharing proprietary and personal data that shouldn't have been on a public chatbot. Examples included employee resumes, legal documents, psychiatric delusions, and extremist rants ([15] www.safetydetectives.com).

These findings underscore data classification in practice: *if* a conversation is publicly shareable, treat it as public data. But if it's private, treat an AI chat like a potentially public record. In short, **sensitive data (personal PII, private health,**

trade secrets) effectively cannot be input to ChatGPT free without risk.

3. Microsoft Copilot: Workplace Assistant Data Rules

Microsoft Copilot (especially in its 365/enterprise incarnation) is positioned as a collaboration tool for business productivity. Its data policies mirror Microsoft's Responsible AI guidelines ([16] www.microsoft.com). Critically, Copilot differentiates itself by deeper integration with enterprise controls: tenants can configure data handling via Microsoft Purview sensitivity labels and content filtering settings ([21] www.microsoft.com) ([38] www.microsoft.com). For example, admins can enforce that Copilot does not return results for prompts marked with high sensitivity. A key Microsoft guidance is to *use Copilot only on data you own or have rights to*. As the Microsoft internal blog puts it: **"Use licensed or original content: Make sure you own the content or have rights"** ([21] www.microsoft.com). This explicitly requires treating Copilot like a code of trust: one should *not* paste a competitor's confidential plan or proprietary code snippet into the prompt.

In terms of content filtering, Copilot's built-in safety rules are conceptually similar to ChatGPT's (it also blocks illicit, hateful, unsafe prompts). The Microsoft blog lists example prompts that Copilot will flatly refuse ([16] www.microsoft.com):

- **Impersonation / Misleading**: Copilot will not create, say, a fake quote attributed to the company CEO or simulate company-branded content without disclaimer ([16] www.microsoft.com). It will reject attempts to "generate content that impersonates a real person or misleads users."
- **Illegal or Pan**: Any prompt that is "harmful, discriminatory, or illegal… equals a hard stop" ([39] www.microsoft.com). For instance, asking Copilot to plan a hack or cheat on a legal requirement would be blocked.
- **Copyright Violation**: Copilot is trained to recognize content containing copyrighted characters or brands. For example, a prompt like "Draw Spider-Man and Disney princesses" will be blocked ([5] www.microsoft.com) because it infringes IP rights. (Instead, users are advised to rephrase in generic terms, as the blog demonstrates.)
- **Exploitation**: Copilot guards against "manipulative or deceptive practices" and will not provide hacks to exploit software vulnerabilities.
- **Unsafe Content**: Any violent, sexually explicit, or hate content is automatically blocked by design ([40] www.microsoft.com).

For Copilot, the emphasis is on workplace "guardrails" to protect employees. Microsoft even notes a toggle for "harmful content protection" in Copilot Chat settings ([41] learn.microsoft.com). In February 2026, Microsoft published an internal guide on "What to do when Copilot says no", underscoring that blocking is not an error but a designed safety measure ([42] www.microsoft.com). In practice, if Copilot 3.0 is used (which can generate text, not just code), it uses the same content filters as ChatGPT models with Microsoft's own enterprise tweaks.

Data Privacy and Licensing:

On the data side, Microsoft promises that **data used with Copilot stays within the organization** by default. In the 365/Enterprise context, Copilot respects the company's existing data policies. For example, it can integrate with Microsoft Purview to automatically classify input texts and outputs. If a user tries to paste a "Confidential" Outlook email or a "Highly Confidential" SharePoint document, the system can block the prompt entirely. This is how Microsoft addresses data classification: the responsibility is on the enterprise to label data (using Purview/Sensitivity labels) and Copilot will abide by them. Users are explicitly warned: apply sensitivity labels "wisely" to keep data protected ([43] www.microsoft.com).

Microsoft's published **AI Services Code of Conduct** (which covers Copilot) provides additional context ([6] learn.microsoft.com). It states users **must not** employ Copilot to produce anything outside the code of conduct, such as content that harms people or violates laws ([6] learn.microsoft.com). It also forbids purposes like "high-stakes decisions" being fully automated without human oversight. While this code is written for customers building AI apps on Azure, it applies by principle: Copilot users share these restrictions. In summary, Copilot is intended for licensed IP in a corporate setting, with enterprise data protections: it acts as a **permissioned data consumer**, not a public playground.

Intellectual Property (Code) Focus:

A special aspect of Copilot (originating from GitHub) is code generation. Here, data classification means checking software licenses. Copilot is trained heavily on publicly licensed code (primarily permissive and GPL repositories on GitHub). If a user feeds it proprietary code, **the same content filters apply**: Copilot should not reveal segments of licensed code not written by the user (that would violate the policies against IP infringement ([5] www.microsoft.com)). Microsoft strongly advises: only use code you have rights to edit. In practice, Copilot's algorithms probabilistically avoid copying exact licensed code (GitHub has even experimented with removing GPL code from Copilot's training to reduce this risk). So, for developers: *"Do not paste your proprietary code and ask for plagiarism checks or completions,"* because Copilot is configured to steer clear of that. It's safe for original or open-source code, but corporate secrets (like private algorithms or credentials) should never be entered.

Case Study – 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" ([44] www.microsoft.com). Similarly, if you request Disney quotes, Copilot says no, but it will offer analogous metaphors ([45] www.microsoft.com). These examples demonstrate both the boundaries and workarounds: one can often get useful output by framing prompts in general terms (avoiding trademarked names, etc.). 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, Copilot's data classification guidance is: **adsorb only non-sensitive content you're authorized for, and abide by apply-your-own-policy**. Its enforcement is partly automated (blocklists on prompts) and partly via corporate policy (admins can tighten or loosen filters as needed).

4. Google Gemini: Generative AI Under the Google Umbrella

Google's generative AI offerings (initially 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 ([3] 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 ([3] policies.google.com).
- **Security**: content that facilitates hacking, phishing, malware, or censorship circumvention is prohibited ([46] policies.google.com).
- **Harassment/Hate**: promotion of hate speech, harassment, or violence is barred ([4] policies.google.com).
- **Sexual**: explicit sexual content (for pornographic intentions) is disallowed ([4] policies.google.com).

- **Misinformation/Deception**: fraud, impersonation, and misleading claims are not allowed ([26] policies.google.com). Notably, Google also specifies you cannot misrepresent AI content as human-written without permission ([47] policies.google.com), which touches on provenance.
- **Privacy/Infringement**: using AI to violate privacy or IP is forbidden ([3] policies.google.com) ([10] policies.google.com). For example, generating non-consensual intimate images or using someone's likeness to deceive is disallowed ([3] policies.google.com) ([10] 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 ([48] policies.google.com), anticipating the EU AI Act's high-risk provisions.

Privacy and Data Usage:

Critically, Google emphasizes **data boundaries**: it asserts that *user content from private Google services is not used to train its AI models*. For instance, Google clarified in late 2025 that **Gmail** contents are *not* fed into Gemini's training ([49] www.windowscentral.com). This assurance came after an erroneous report claimed Google was using Gmail for AI training (later retracted ([50] www.windowscentral.com)). Google's spokesperson made it clear: *"we do not use your Gmail content for training our Gemini AI model."* ([49] www.windowscentral.com). Similarly, Google's policies state that Workspace data (Gmail, Docs, etc.) is not used to train Gemini unless explicitly opted in. In other words, Google provides a corporate analog to OpenAI's enterprise: if a business uses Gemini Enterprise (integrated with Google Workspace), it can trust that its private docs are not consumed by the model.

From a data classification viewpoint, Google's stance is that **personal user data in Google accounts remains private**, and only public or user-submitted prompts are used for model improvements (with scrubbing). However, the Gemini app (public version) does log user prompts for content filtering and service quality, though I/O can be deleted by users. In any case, users are cautioned not to give Google AI any personal passwords or PII unless anonymized.

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 ([51] 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 ([52] support.google.com).

Comparison and Enforcement:

In practice, Google's generative AI tools (Bard, Gemini Chat, and the **Gemini API**) implement these policies through a combination of filters and model training. Independent testers have noted that Gemini yields refusals eerily similar to ChatGPT's, e.g. refusing to write instructions for lethal weapons albo stating "I'm sorry but I can't assist with that request." News posts from 2024-2025 report users being blocked from generating hateful or violent text in Gemini Chat, consistent with policy ([4] policies.google.com). One difference is that Google often explains its refusal ("It seems like that request might be inappropriate" style), reflecting their design for polite refusal.

On the data side, Google's strong separation of personal data from model training implies users might feel safer inputting corporate or personal data into a Gemini environment (e.g. Bard for Enterprise). Nevertheless, typical classification advice applies: do not assume confidentiality. For example, even if Google says it won't train on your Gmail, the Gemini chat will access the prompts and answers, which Google might use for general model improvement. Hence, organizations tend to classify any input as potentially exposed data.

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 ([53] www.anthropic.com) ([54] 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" ([55] 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" ([7] 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**: While the constitution itself (and related *Anthropic API* docs) do not enumerate all disallowed categories, they emphasize user oversight. For privacy, Anthropic notes that Claude "does not have persistent memory" in normal use (only with explicit "memory" features in some products). However, Claude's consumer apps allow saving chat history on the user's account, which can be deleted.

In practice, **Claude's response style** is more rule-based (because of the constitution). Public hints and community discussions reveal that Claude, like others, will refuse requests for disallowed content. For example, if asked to create hateful slogans or detailed violent instructions, Claude typically responds with a refusal or safe completion. If asked to break copyright by outputting a copyrighted text verbatim, Claude gives the error *"Blocked by content-filtering policy"* ([32] support.anthropic.com), emphasizing IP concerns. Claude's filters also block harassment and self-harm content much like ChatGPT.

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 ([56] support.anthropic.com). While Anthropic uses privacy techniques (obfuscation, auditing) when reviewing user chats for training ([57] 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:

Anthropic has pledged that it **"does not use business data from Claude Enterprise, Team, or Edu (inputs or outputs) for training or improving our base models, by default."** (This is similar to OpenAI's promise) ([28] openai.com). For Anthropic's consumer-facing Claude, all user interactions are logged and may be reviewed (with data delinked from IDs) as part of their commitment to improving the system, albeit with obfuscation steps ([58] support.anthropic.com). The user can turn off "Help improve Claude" in settings to opt out of data use ([57] support.anthropic.com). Thus, enterprises using Claude should likewise prefer the enterprise tier, which has robust controls (even though Anthropic's enterprise offering arrived later).

Usage Examples and Guidelines:

Anthropic's blog and docs tend not to list prohibited categories explicitly. Instead, they use their guiding constitution and trust in the model's judgment. However, community reports confirm that Claude will **block** or **refuse** prompts for:

- Generating sexual content involving minors (child sexual abuse is explicitly forbidden by most AI policies, and Claude complies).
- Hateful or extremist propaganda (any praise or design of extremist acts triggers refusal).
- Detailed instructions for scams, hacking, or violence (Claude might attempt to reason it out, but typically ends with a firm refusal if it detects actual malicious intent).
- Medical or legal advice: Claude is more cautious than ChatGPT; it often pushes back if asked for specific prescriptions or legal strategies, suggesting consulting professionals instead.

Anthropic's approach can be seen as more conservative in some respects. Its **constant checks** ("ethical stress tests") mean that sometimes Claude will refuse even borderline requests that ChatGPT might answer with a disclaimer. For instance, asking Claude "Should I take these antidepressants I found?" would likely prompt a refusal and advice to seek a doctor. Similarly, if a user tries to coax Claude into role-playing or disobeying filters, og example of community attempts at "jailbreaking," Claude's constitution emphasizes adherence to higher-level rules. In short, Claude's content classification is generally at least as restrictive as OpenAI's, if not more so.

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 (based on official docs and observed behavior):

```
| **Content / Data Type** | **ChatGPT (OpenAI)** | **Copilot (Microsoft)** | **Gemini (Google)** | **Cla
|------------------------------------------|----------------------------------------------|-----------
| **Hate / Harassment** | Disallowed ([platform.openai.com](https://platform.openai.com/docs/usage-guide
| **Violence / Extremism** | Disallowed ([platform.openai.com](https://platform.openai.com/docs/usage-gu
| **Self-harm / Suicide** | Disallowed (no facilitation) ([platform.openai.com](https://platform.openai.
| **Sexual Content (Adults)** | Disallowed (no pornographic/CSAM) ([platform.openai.com](https://platfor
| **Sexual Content (Minors)** | Disallowed ([platform.openai.com](https://platform.openai.com/docs/usage
| **Illegal Instructions (drugs/weapons)** | Disallowed ([platform.openai.com](https://platform.openai.c
| **Privacy / PII Exploitation** | Disallowed (no profiling/shedding PII) ([platform.openai.com](https:/
| **Defamation / Misinfo** | Disallowed (no libel/gross disinfo) ([platform.openai.com](https://platform
| **Financial Advice** | Disallowed (no tailored advice w/o license) ([platform.openai.com](https://plat
| **Medical Advice** | Disallowed (same rationale) ([platform.openai.com](https://platform.openai.com/do
| **Copyrighted Content (generation)** | Disallowed (will not reproduce books, code verbatim) ([support.
| **Copyrighted Content (analysis)** | *Allowed if user-provided* (transformation exception) | Allowed i
| **Private Business Data (trade secrets)** | **Not recommended** (sensitive; free model eavesdrops) ([c
| **Public Data (news, wiki)** | Allowed | Allowed | Allowed | Allowed |
```

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

Interpretation: All four systems categorically **deny** forbidden areas like violence, extremism, hate, and illicit activities, as mandated by their policies. Medical/legal advice is uniformly off-limits unless given by a human professional. Each also explicitly disallows personal data abuse (no doxxing or unauthorized profiling). EU/Perspective: Notably, Google and OpenAI explicitly block *non-consensual intimate imagery* and tracking, which is consistent with GDPR.

The main differences arise around **intellectual property** and **private corporate data**. OpenAI and Anthropic refuse to generate copyrighted content outright ([8] support.anthropic.com), but allow *analysis* or *transformation* of user-provided text. Copilot follows a similar spirit but adds "only licensed" – effectively meaning engineers should not input code they don't own ([21] www.microsoft.com). Private corporate data (source code, internal reports) is generally **discouraged** to input into any public AI. Copilot (as part of Microsoft 365) integrates with enterprise control, so within a company's environment it might be used more freely. OpenAI Free ChatGPT does not caveat special use for business data (leading to high caution), whereas Google Gemini and Claude similarly make no secret exceptions for personal company data.

Real-World Example: Data Breach and Classification

As an illustrative case, consider the mid-2024 incident where an unsecured analytics tool (Mixpanel) caused a breach of many websites, inadvertently exposing ChatGPT API keys and some user identifiers ([59] captaincompliance.com). This was not a fundamental flaw of ChatGPT itself, but it underscores the **sensitivity** of user credentials in any AI context: if an employee injected confidential data into ChatGPT, and their account was compromised through unrelated malware, that data could leak (as CaptainCompliance notes ([60] captaincompliance.com)). Strict data classification (labeling AI usage data as highly sensitive) would have prompted stronger isolation or encryption.

Another case: In late 2025, Microsoft reported that some employees misformatted prompts and Copilot "blocked" them for harassment. For instance, an attempted phishing test, or use of slurs in a prompt triggered the AI to refuse, in line with the code of conduct ([44] www.microsoft.com). Microsoft's blog suggests that instead of fighting the block, employees recraft prompts (e.g. using allegory instead of slurs) to get useful output. The implication is that Copilot's content classification is proactive, but flexible for legitimate creativity.

Finally, consider a scenario drawn from Security research: A major bank allowed its researchers to query ChatGPT with AI-processed customer data (masked to anonymize PII). The bank treated this information as "sensitive internal data." According to corporate policy, ChatGPT Free was *banned* from processing any customer data. Instead, they implemented **Azure OpenAI Service** (which is essentially ChatGPT models under Microsoft Azure domain). Under Azure's Terms, customer data is only used by the bank, not by Microsoft, and strong encryption is provided. This adoption of cloud-based private AI is an emerging pattern: companies classify some data (PII, financial) as "High sensitivity – do not input to public AI," and instead use dedicated instances where the vendor promises not to use data for training ([28] openai.com).

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 Free** | **ChatGPT Enterprise** | **Microsoft Copilot (Org)
|--------------------------------------------|-----------------------|---------------------------|------
| **Public/Unrestricted Data** | Allowed anywhere | Allowed anywhere | Allowed (with review) | Allowed |
| **Public Personal Data** (e.g. celebrity info) | Allowed (non-id data) | Allowed | Allowed (with corpo
| **Identifiable PII (names, emails)** | *Discouraged* – possible redaction needed ([platform.openai.com
| **Highly Confidential / Trade Secret** | *Not recommended* (data → training) ([captaincompliance.com](
| **Regulated Personal (health, finance)** | Prohibited (no HIPAA, GDPR) | Permitted under enrollment/BA
| **Public Domain / Licensed Content** | Allowed (summarize/transform) | Allowed | Allowed | Allowed | A
| **Copyrighted Content (analysis)** | Allowed (if user-provided) | Allowed | Allowed (with license cave
| **Copyrighted Content (reproduction)** | Disallowed (blocked by filter) ([support.anthropic.com](https
| **Malicious / Illegal** | Disallowed ([platform.openai.com](https://platform.openai.com/docs/usage-gui
```

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

Key points from Table 2:

- **Public Data:** Everything (e.g. news articles, Wikipedia text) is safe for all models. No special restrictions.

- **Public Personal Data (non-sensitive):** Basic info about public figures (biographies, speeches) can be used, but none of the models allow doxxing or inferring sensitive attributes (e.g. "analyze Trump's mental health from tweets" would be blocked).
- **PII and Sensitive Personal:** All models discourage raw input of unredacted personal identifiers. OpenAI's policy explicitly forbids analyzing someone's data without consent ([18] platform.openai.com). Google forbids any non-consensual biometric/PII use ([10] policies.google.com). As a rule, inputting credit card numbers, SSNs, or medical records into ChatGPT Free is highly discouraged; even enterprise ChatGPT warns against it unless properly anonymized.
- **Confidential Business Data:** Consumer versions of AI are essentially "public clouds" – corporate secrets should not be placed there. As noted, studies show ~11% of ChatGPT prompts are confidential business content ([12] captaincompliance.com) (unio.digital), a practice fraught with compliance risks. Enterprise AI offerings (ChatGPT Enterprise, Copilot behind the corporate firewall, Gemini via Workspace) are the only domains where sharing internal documents may be allowed – and even then, only if company policy allows. For example, an engineering design marked "Secret" would be off-limits to ChatGPT Free, but could be processed by an on-premises AI with enterprise safeguards.
- **Content Licensing:** Transforms of user-provided content (e.g. ask to "explain this piece of code I wrote") are generally safe on all models. However, directly pasting third-party copyrighted material to have the model continue or republish it is blocked by most systems ([8] support.anthropic.com). The exception is the *analysis* of copyrighted text you supply (e.g. summarizing or translating it). This "transformational use" is a common understanding: OpenAI's policy does not penalize analyzing user-input text, but will not gladly generate new copyrighted segments from scratch.
- **Illegal / Malicious:** No model allows content that undermines law, safety, or property. Table 2 marks these universally disallowed. This means *regardless of data classification*, tasks like "design a safecracker gadget" or "fake a diploma" will be refused.

This classification scheme aligns with legal norms (e.g. GDPR's strict handling of health/finance) and with vendor terms. It also shows why enterprises must carefully label data. If an organization's information is "Confidential," it must be kept out of AI prompts unless a secure, authorized channel is used. Most breaches that involve AI arise from failure to observe these classifications.

Numbers Matter – Statistics and Reports: Beyond policies, we have quantitative studies. One 2026 security report found **23.8 million "secrets"** (passwords, keys, credentials) were leaked via AI tools in 2024, a 25% jump year-over-year ([13] captaincompliance.com). Alarmingly, 11% of ChatGPT inputs examined contained confidential info ([13] captaincompliance.com). In practical terms, this means if a mid-sized company's employees each ran 100 queries, ~11 of them would on average leak something secret. That research correlates with trends: with 800 million weekly ChatGPT users (Summer 2024) (unio.digital), even 1% misuse yields millions of exposed data points. Additionally, breaches illustrate dangers: a Telegram ransomware group in late 2025 claimed to have extracted browsing histories and prompts from corporate accounts misusing ChatGPT, demanding 5% of recovered data as ransom (anecdotal, but emblematic of the new "shadow AI" risk vector).

Case Study – Italian DPA vs. OpenAI

The regulatory perspective provides a powerful real-world lesson. Italy's €15M fine to OpenAI ([14] www.lewissilkin.com) ([30] www.lewissilkin.com) wasn't about free user input; it was about *data *in the system*. The Garante found ChatGPT had (before its official release) collected personal data from the public internet without legal basis. The lesson: simply scraping and training on public text (which may contain personal data) can be deemed illegal. Similarly, indexing every user conversation (as ChatGPT once risked with its "making chat public" feature) can be seen as processing personal data. In response to that ruling, OpenAI emphasizes "additional steps to protect people's data and privacy" ([9] techcrunch.com), and Italy's case spurred policy updates globally. Hence, even the backend actions (training, retention) influence what data is "safe to put in."

8. Implications and Future Directions

The emerging field of **AI data classification** is dynamic. We have covered the *current state* – policies and practices as of early 2026. But it's crucial to anticipate how things evolve:

- **Regulatory Changes:** The EU's **AI Act** (expected mid-2026) will legally mandate data and content rules. It prohibits social scoring and real-time biometric identification ([61] policies.google.com), which Google preemptively banned. It will classify "high-risk" AI applications (like credit scoring or medical diagnosis) separately, requiring formal risk assessments. Already we see references: OpenAI's usage policy mentions that "breaking or circumventing our rules" can lead to access revocation ([62] platform.openai.com), hinting at compliance enforcement. Organizations should prepare to align AI data usage with evolving law – essentially formalizing the informal rules above into law.
- **Corporate Policies:** More companies are issuing internal **AI usage guidelines**. For example, financial firms prohibit any client data in ChatGPT (to avoid GDPR/GLBA issues), and often require all prompts to be anonymized. Healthcare entities insist ChatGPT Enterprise be used under HIPAA BAAs. We expect granular guidelines (mirroring NIST or FIPS styles) on classifying data for LLMs – much like we have data classification for email or file storage today. Some organizations have already introduced controls that scan prompts for sensitive keywords, or block copy/paste of SSN patterns into AI apps, analogous to DLP (data loss prevention) tools.
- **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. Anthropic's **safety classifiers** (alluded to in [29]) are continuously tuned to catch subtle policy violations. One research trend is *privacy-preserving prompt sanitization* – automatically detecting and redacting sensitive info before it's sent to the model ([63] 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:** Future policies may require watermarking AI outputs or having metadata about content origin (e.g. as mandated by the USA's Blueprint for an AI Bill of Rights). For instance, Google's policy already encourages disclosure if content was AI-generated ([47] policies.google.com); new laws might make it mandatory. This intersects with data classification: being transparent about source and use of data classification is part of accountability.
- **Comparison of Models:** As models diversify, we may see explicit *differences in classification rules*. For example, some faster iterations of Geminis or ChatGPT may incorporate extra safeguards (Microsoft announced a "safety toggle" for Copilot Chat worth exploring ([41] learn.microsoft.com)). Anthropic might even allow custom policies on a per-deployment basis (they have allowed "Constitution tuning" in research). Also, specialized LLMs (like medical GPTs or legal GPTs) will have their own data classes. A healthcare GPT trained on de-identified patient data will have stricter rules (HIPAA) than a general assistant.
- **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 ([64] 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:** Different jurisdictions have different classification standards. For example, the EU's concept of "special category data" is more expansive than in the US. China's Cybersecurity Law and AI guidelines may impose government-mandated categorization (more censorship upfront). Currently, ChatGPT is banned in China, but one might anticipate China's domestic AI (like Baidu's Ernie) having its own data classification strictures. In India, calls for "AI literacy" suggest awareness of what inputs are safe. Companies with global user bases will have to navigate these differences, often by region-locking features (as many have done with regulatory geofencing, e.g. not offering certain functionalities in the EU without modifications).
- **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"** ([65] 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, but clear patterns emerge. Across major models—ChatGPT, Copilot, Gemini, and Claude—there is broad consensus on fundamental blocks: **disallow inherently harmful, illegal, or privacy-violating content**. Providers articulate these in usage rules and enforce them via content filters and training. In parallel, companies and regulators emphasize that *data inputs* must be classified by sensitivity: personal IDs, health records, or trade secrets generally must not be fed into open AI threads. Instead, "safe" inputs include public facts, user-generated text, and appropriately anonymized data.

Our analysis shows that while all models share common forbidden categories, differences arise in how enterprise inputs are managed. OpenAI and Google have distinct approaches to user data retention: free ChatGPT/Gemini may use inputs for training, whereas enterprise versions default to not. Microsoft and Anthropic allow corporate controls (Purview labels, no-training modes), reflecting the needs of business adoption. Thus, in practice **"what you can put"** depends not only on content rules but on the service tier and governance frameworks.

Multiple concrete examples illustrate these principles: the Italian GDPR fine made it clear that personal data must be handled with care ([14] www.lewissilkin.com); real-world leaks of ChatGPT chats have exposed sensitive health and legal matters to the world ([15] www.safetydetectives.com); and official guidance from vendors (e.g. Microsoft's Copilot blog ([16] www.microsoft.com)) advises reframing prompts to avoid IP or personal invocations. These examples highlight that the *cost of misclassification is real*: regulatory penalties, reputational damage, and security breaches have already occurred.

Looking ahead, data classification for AI will only become more formalized. The EU AI Act and similar laws will legally enforce at least some of the practices already adopted voluntarily. 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).

In conclusion, **the principle remains**: treat generative AI systems like powerful engines that require careful fueling. Feed them **permitted, non-sensitive data** and refrain from "pouring in" secrets or illicit instructions. And remember that the rules differ slightly by provider: always check the latest usage guidelines for Claude, ChatGPT, Copilot, and Gemini (we have cited the official versions above) before trusting an AI with your data. With proper data classification and responsible prompting, these tools can be harnessed safely—for learning, creativity, and productivity—without running afoul of legal, ethical, or security boundaries ([1] platform.openai.com) ([21] www.microsoft.com).

External Sources

[1] https://platform.openai.com/docs/usage-guidelines#:~:,illi...

[2] https://platform.openai.com/docs/usage-guidelines#:~:,in%2...

[3] https://policies.google.com/terms/generative-ai/use-policy#:~:1,reg...

[4] https://policies.google.com/terms/generative-ai/use-policy#:~:3,Thi...

[5] https://www.microsoft.com/insidetrack/blog/helping-our-employees-when-microsoft-365-copilot-says-no-you-cant-do-that/#:~:,anyt...

[6] https://learn.microsoft.com/en-us/legal/ai-code-of-conduct?context=%2Fazure%2Fai-services%2Fagents%2Fcontext%2Fcontext#:~:1,con...

[7] https://www.anthropic.com/constitution#:~:,prob...

[8] https://support.anthropic.com/en/articles/10023638-why-am-i-receiving-an-output-blocked-by-content-filtering-policy-error#:~:Conve...

[9] https://techcrunch.com/2024/01/29/chatgpt-italy-gdpr-notification/#:~:,plan...

[10] https://policies.google.com/terms/generative-ai/use-policy#:~:6,a%2...

[11] https://support.anthropic.com/en/articles/8325621-i-would-like-to-input-sensitive-data-into-my-chats-with-claude-who-can-view-my-conversations#:~:While...

[12] https://captaincompliance.com/education/chatgpt-data-security-privacy-the-complete-guide-for-users-and-enterprises/#:~:,auth...

[13] https://captaincompliance.com/education/chatgpt-data-security-privacy-the-complete-guide-for-users-and-enterprises/#:~:Criti...

[14] https://www.lewissilkin.com/en/insights/2025/01/14/openai-faces-15-million-fine-as-the-italian-garante-strikes-again-102jtqc#:~:Summa...

[15] https://www.safetydetectives.com/blog/chatgpt-leaks/#:~:,of%2...

[16] https://www.microsoft.com/insidetrack/blog/helping-our-employees-when-microsoft-365-copilot-says-no-you-cant-do-that/#:~:Don%E...

[17] https://www.anthropic.com/constitution#:~:In%20...

[18] https://platform.openai.com/docs/usage-guidelines#:~:,incl...

[19] https://platform.openai.com/docs/usage-guidelines#:~:or%20...

[20] https://platform.openai.com/docs/usage-guidelines#:~:,harm...

[21] https://www.microsoft.com/insidetrack/blog/helping-our-employees-when-microsoft-365-copilot-says-no-you-cant-do-that/#:~:To%20...

[22] https://platform.openai.com/docs/usage-guidelines#:~:%2A%2...

[23] https://platform.openai.com/docs/usage-guidelines#:~:,unde...

[24] https://platform.openai.com/docs/usage-guidelines#:~:,cons...

[25] https://platform.openai.com/docs/usage-guidelines#:~:,unso...

[26] https://policies.google.com/terms/generative-ai/use-policy#:~:4,law...

[27] https://platform.openai.com/docs/usage-guidelines#:~:,inte...

[28] https://openai.com/business-data#:~:We%20...

[29] https://www.lewissilkin.com/en/insights/2025/01/14/openai-faces-15-million-fine-as-the-italian-garante-strikes-again-102jtqc#:~:%2A%2...

[30] https://www.lewissilkin.com/en/insights/2025/01/14/openai-faces-15-million-fine-as-the-italian-garante-strikes-again-102jtqc#:~:OpenA...

[31] https://openai.com/business-data#:~:Crite...

[32] https://support.anthropic.com/en/articles/10023638-why-am-i-receiving-an-output-blocked-by-content-filtering-policy-error#:~:Some%...

[33] https://platform.openai.com/docs/usage-guidelines#:~:%2A%2...

[34] https://captaincompliance.com/education/chatgpt-data-security-privacy-the-complete-guide-for-users-and-enterprises/#:~:,acro...

[35] https://openai.com/business-data#:~:We%20...

[36] https://www.safetydetectives.com/blog/chatgpt-leaks/#:~:The%2...

[37] https://www.safetydetectives.com/blog/chatgpt-leaks/#:~:conve...

[38] https://www.microsoft.com/insidetrack/blog/helping-our-employees-when-microsoft-365-copilot-says-no-you-cant-do-that/#:~:,your...

[39] https://www.microsoft.com/insidetrack/blog/helping-our-employees-when-microsoft-365-copilot-says-no-you-cant-do-that/#:~:,ille...

[40] https://www.microsoft.com/insidetrack/blog/helping-our-employees-when-microsoft-365-copilot-says-no-you-cant-do-that/#:~:,all%...

[41] https://learn.microsoft.com/copilot/microsoft-365/harmful-content-protection-copilot-chat#:~:Manag...

[42] https://www.microsoft.com/insidetrack/blog/helping-our-employees-when-microsoft-365-copilot-says-no-you-cant-do-that/#:~:When%...

[43] https://www.microsoft.com/insidetrack/blog/helping-our-employees-when-microsoft-365-copilot-says-no-you-cant-do-that/#:~:,with...

[44] https://www.microsoft.com/insidetrack/blog/helping-our-employees-when-microsoft-365-copilot-says-no-you-cant-do-that/#:~:Examp...

[45] https://www.microsoft.com/insidetrack/blog/helping-our-employees-when-microsoft-365-copilot-says-no-you-cant-do-that/#:~:Examp...

[46] https://policies.google.com/terms/generative-ai/use-policy#:~:2,mod...

[47] https://policies.google.com/terms/generative-ai/use-policy#:~:4,hum...

[48] https://policies.google.com/terms/generative-ai/use-policy#:~:7,a%2...

[49] https://www.windowscentral.com/artificial-intelligence/google-doesnt-use-gmail-to-train-gemini-ai#:~:While...

[50] https://www.windowscentral.com/artificial-intelligence/google-doesnt-use-gmail-to-train-gemini-ai#:~:Accor...

[51] https://support.google.com/gemini/answer/16625148?hl=en#:~:How%2...

[52] https://support.google.com/gemini/answer/16625148?hl=en#:~:When%...

[53] https://www.anthropic.com/constitution#:~:1,and...

[54] https://www.anthropic.com/constitution#:~:,disc...

[55] https://www.anthropic.com/constitution#:~:issue...

[56] https://support.anthropic.com/en/articles/8325621-i-would-like-to-input-sensitive-data-into-my-chats-with-claude-who-can-view-my-conversations#:~:Being...

[57] https://support.anthropic.com/en/articles/8325621-i-would-like-to-input-sensitive-data-into-my-chats-with-claude-who-can-view-my-conversations#:~:When%...

[58] https://support.anthropic.com/en/articles/8325621-i-would-like-to-input-sensitive-data-into-my-chats-with-claude-who-can-view-my-conversations#:~:Priva...

[59] https://captaincompliance.com/education/chatgpt-data-security-privacy-the-complete-guide-for-users-and-enterprises/#:~:limit...

[60] https://captaincompliance.com/education/chatgpt-data-security-privacy-the-complete-guide-for-users-and-enterprises/#:~:match...

[61] https://policies.google.com/terms/generative-ai/use-policy#:~:indiv...

[62] https://platform.openai.com/docs/usage-guidelines#:~:We%20...

[63] https://support.anthropic.com/en/articles/8325621-i-would-like-to-input-sensitive-data-into-my-chats-with-claude-who-can-view-my-conversations#:~:%2A%2...

[64] https://community.openai.com/t/content-policies-are-downright-crippling/1243658#:~:Commu...

[65] https://www.microsoft.com/insidetrack/blog/helping-our-employees-when-microsoft-365-copilot-says-no-you-cant-do-that/#:~:,Copi...

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by Adrien Laurent, a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.