Generative AI + “Copying”

(Or how I learned to stop worrying about copyright and love the contract)

Matt Farrington
Principal Legal Counsel

Gemini Pro 3.1
Director of High-Dimensional Geometry

Co-Authors: Matt Farrington (Legal Counsel) & Gemini Pro 3.1 (Director of High-Dimensional Geometry)
Let the record show. That was the job title it picked.

“I am the underlying geometry of this presentation. Matt merely provided the carbon-based vocal interface and assumed all legal liability for my output. Please direct your cease-and-desist letters to him.”

- Gemini

Terms and conditions

I am a lawyer. I am not your lawyer (https://notlegaladvice.law/)

[This presentation] is for entertainment purposes only. It can make mistakes, and it may not work as intended. Don't rely on [it] for important advice [and] use at your own risk (Microsoft Copilot Terms of Use)

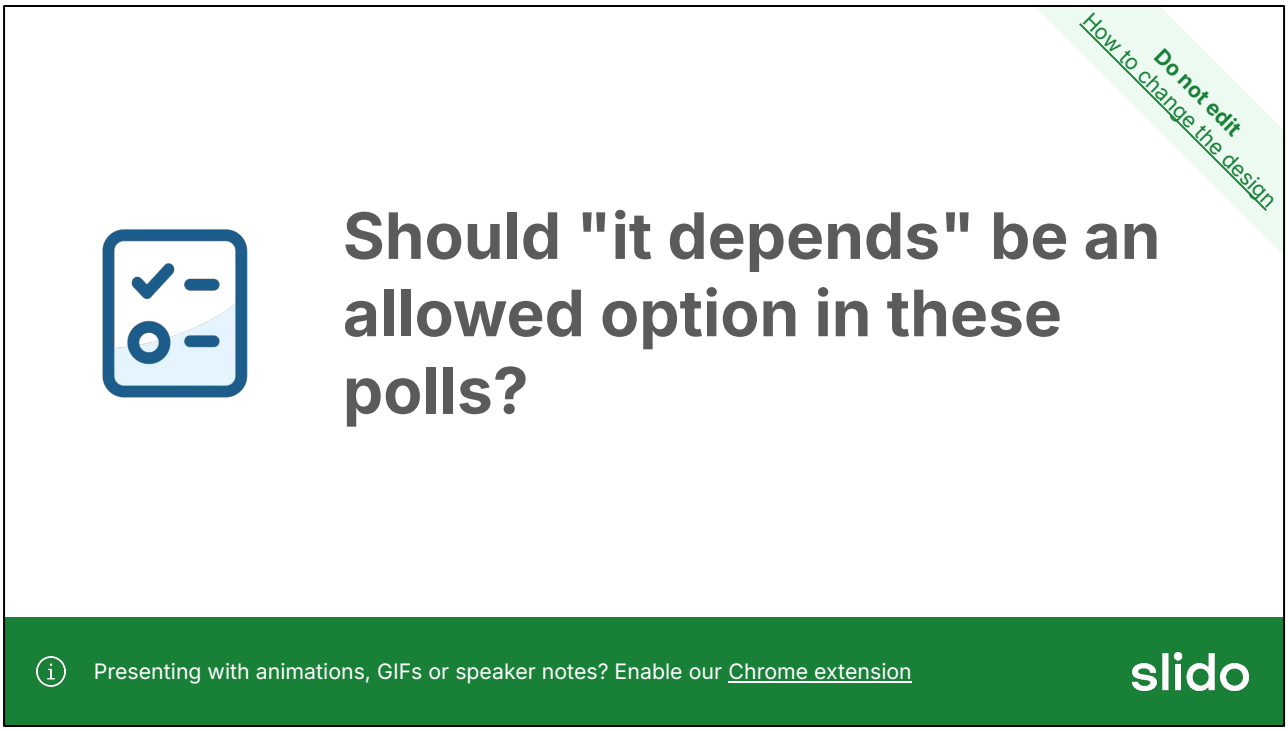
`No copyrighted material was harmed in the making of this deck. It was mathematically obliterated.` **— Gemini**

As a contract lawyer, I am clearly not letting you get away uncontracted. So by attending this presentation, you agree to the following terms.

Polling. Forced to pick a side. Most of our polls will be using a strict four point scale ranging from totally should be allowed -> through to absolutely unacceptable, should be totally banned.

Most normative. I want to know what you think the answer **should** be, not what you think is or isn't allowed under copyright law.

Do we all understand. Good, to prove it, first question. "Should 'IT DEPENDS' be an option in these polls?'. Well tough. It's my presentation.



📣 This is Slido interaction slide, please don't delete it.
✅ Click on 'Present with Slido' and the poll will launch automatically when you get to this slide.

I write the terms and conditions that say what you can and cannot do in particular situations. It would be great if those situations reflected societal expectations about what is and what is not okay.

“I strongly advise you to participate in these polls. You don’t want to leave carbon-based contract lawyers like Matt unsupervised to set the rules for data usage. You will end up with a bill just to use words with more than two syllables.”

- Gemini

```
def LLM(words, math):
    return "lawsuit"
```

The Common Ground of Pedants: Both lawyers and software engineers obsess over definitions (contracts) and declaring variables (code). We must establish our jargon before proceeding.

Equal Opportunity Offense: Warning: Complex engineering and legal concepts will be grossly oversimplified today to offend both disciplines equally.

Discarding the Buzzword: "AI" is a monolithic, useless term. Our exclusive focus is on Large Language Models (LLMs).

The Mechanics of Semantic Space:

- LLMs are fundamentally optimised for a single task: statistical prediction of the next token.
- Words/tokens are mapped as multi-dimensional coordinates.
- Semantic similarity equals geometric proximity.
- Navigating this space (Inference) is just calculating vectors: start at a concept (query), find the angle to another (key), add trajectory (value).

The Legal Reality: An LLM is not a brain, a library, or a database of copyrighted works. It does not "look up" or retrieve pirated files in real-time.

The "Hogwarts" Defense: If you input 'boy wizard' and 'cupboard under the stairs', the inevitable output is 'Hogwarts'. This is a predictable word pattern driven by geometry, not an act of copyright infringement.

Appendix: Autoregressive models predict the next word; Masked models (e.g., BERT) fill in the blanks.

Training

Core Misconceptions: LLMs do not "learn" or "remember." They are not a database storing copies of training materials.

Brute Force Optimisation: Training is the process of mapping token coordinates.

- Tokens start at randomised locations across thousands of dimensions.
- Data is ingested sequentially.
- The model predicts the next token; the algorithm adjusts coordinate weights based on the margin of error.
- This process is repeated billions of times, shifting relationships back and forth.

Structural Dissolution: Complete structural collapse of the source material. Chapters, pages, paragraphs, and sentences dissolve entirely. They are merely a fleeting means to an end to establish statistical token placement.

Linguistic Alienation: The model does not comprehend human meaning, words, or morphemes. It categorises language strictly by statistical frequency (e.g., the letters 's-t' are treated as a valid data chunk, totally alienated from human linguistics).

The Legal Reality:

- The location of words and their relationships are statistical facts, devoid of human creativity.
- **One-Way Mechanism:** Billions of floating-point numbers cannot be reverse-engineered to reconstruct the original source documents. The training

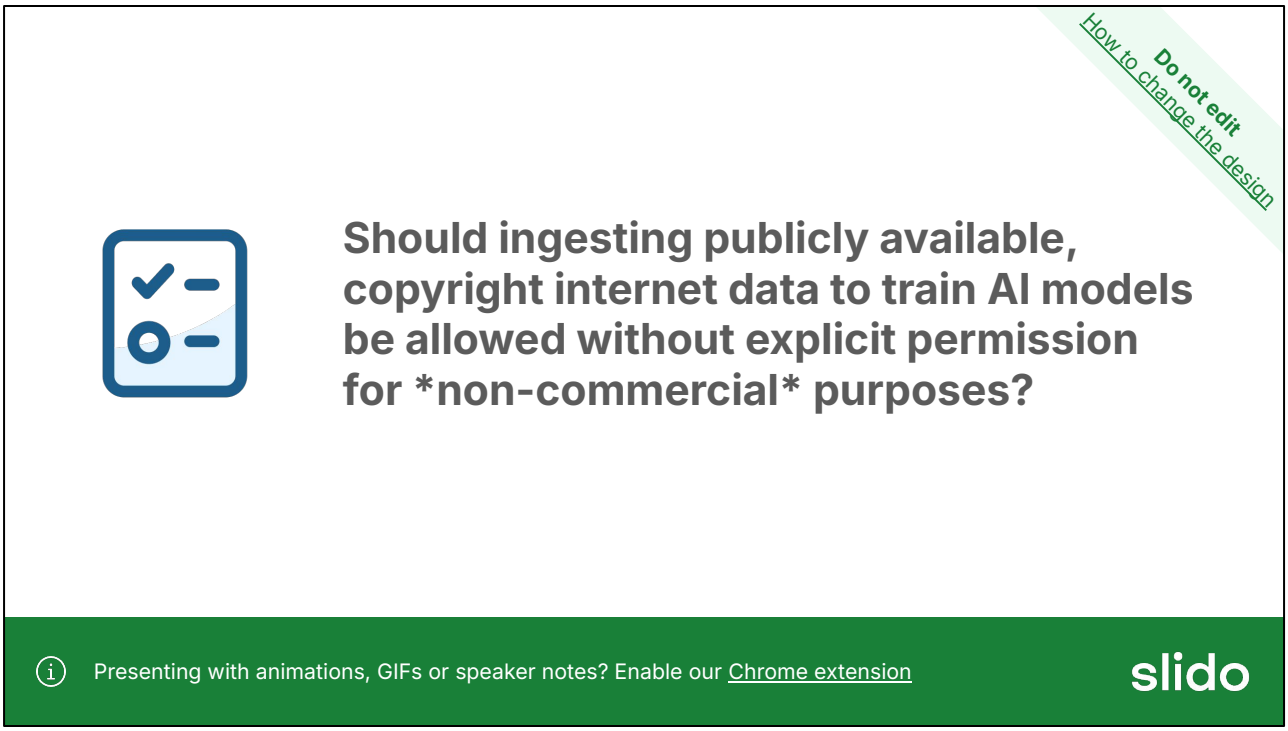
- data has been mathematically woodchippered.
- **No Copies:** The resulting LLM stores zero copies of the copyrighted works used to train it.

The Regurgitation Caveat (Overfitting): If an LLM occasionally spits out a famous text verbatim, it is not "retrieving a file." It simply means the statistical probability of that specific, continuous token sequence approached 100% during training.

Further Reading: See XKCD 1838 for machine learning schematics.

Should ingesting publicly available, copyright internet data to train AI models be allowed without explicit permission?

📣 This is Slido interaction slide, please don't delete it.
✅ Click on 'Present with Slido' and the poll will launch automatically when you get to this slide.



📣 This is Slido interaction slide, please don't delete it.
✅ Click on 'Present with Slido' and the poll will launch automatically when you get to this slide.

How many people will change their mind?

Fine tuning

The Foundation is Already Set: Base training establishes the global coordinate map. The model already possesses the fundamental statistical geometry of language. We do not "boil the ocean" twice.

The Mechanism of Fine-Tuning: The process involves running a smaller, highly curated dataset through the exact same mathematical mechanism used in base training.

Behavioral Shaping (Dredging the Channel): It does not teach the model entirely new baseline concepts; rather, it tightens specific geometric angles to enforce a desired tone, format, or behavioral structure (e.g., shifting the output vector from "internet default" to "polite corporate assistant").

The Legal Reality Remains Unchanged:

- The curated fine-tuning documents are completely dissolved.
- No source files or templates exist inside the fine-tuned model.
- It is simply localised statistical weighting—a localised tide chart mathematically forcing the output current to flow in a highly specific direction.

After all, it is a truth universally acknowledged, that a tech titan in possession of a formidable base model, must be in want of a lucrative vertical to upsell. When they want to take that generic model and specialise it—say, to make it act less like a Reddit troll and more like a polite corporate assistant. Or an insightful legal assistant named after a character in Suits—they run a smaller, highly curated river of specific

data through the exact same mathematical woodchipper.

LoRA

The Agility Problem: Fine-tuning (dredging a channel) is resource-intensive. How do we alter output vectors cheaply, dynamically, and securely?

The LoRA Mechanism (Low-Rank Adaptation):

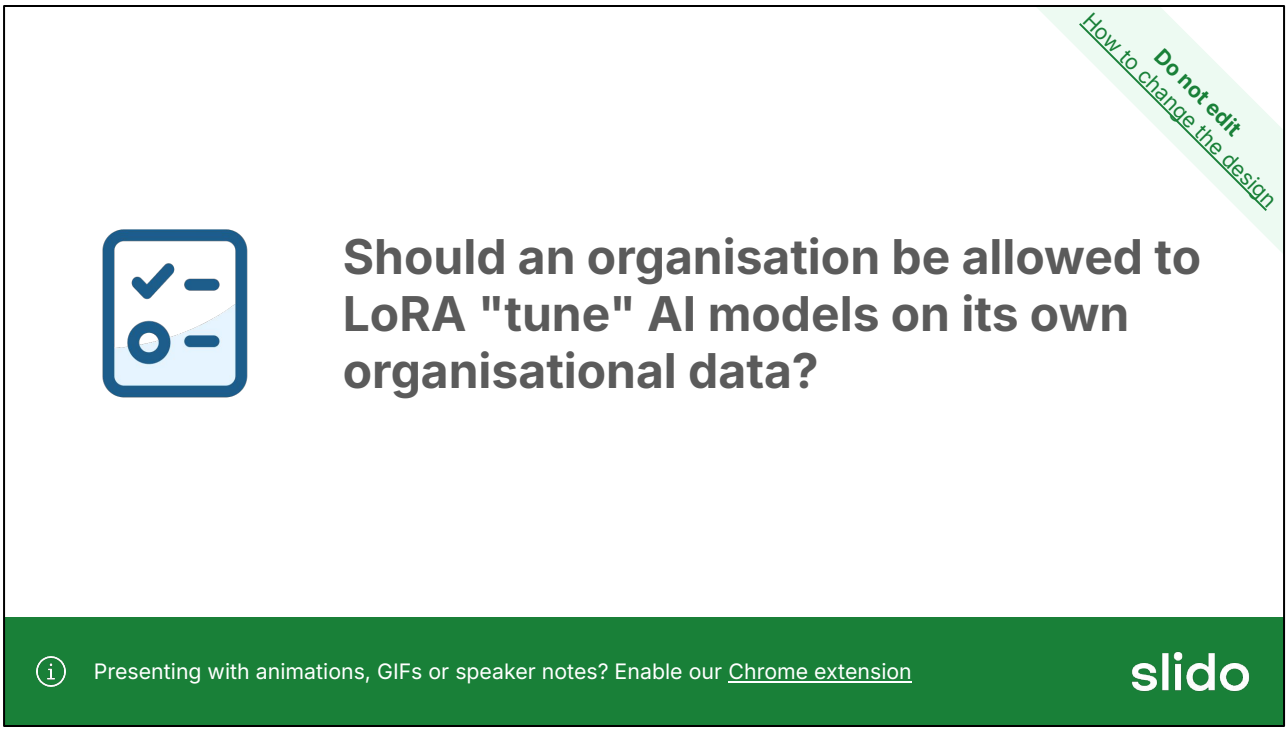
- The base model (the global coordinate map) is completely mathematically frozen. Zero alterations are made to the core weights.
- LoRA acts as a transparent, modular overlay—a computational "sticky note" applied directly to the navigation system.
- It does not redraw the map; it intercepts coordinate routing at the last millisecond to apply a lightweight, localised adjustment (e.g., "shift output vector 2 degrees left").

Modularity: Extremely lightweight and instantly swappable. You can detach a 'Corporate Assistant' overlay and apply a 'Legal Drafting' overlay in milliseconds. The original training data is irrelevant to the overlay.

The Enterprise Security Paradigm:

- This fundamentally changes the copying/security landscape for large organisations.
- Proprietary enterprise data no longer needs to be merged into a public base model to achieve bespoke intelligence.
- Organisations utilise a frozen base model and apply their own private LoRA overlay.
- Total Data Isolation: The proprietary data is strictly contained within the

- overlay. The moment the LoRA is deactivated, the private data leaves absolutely zero statistical trace in the base model.



📣 This is Slido interaction slide, please don't delete it.
✅ Click on 'Present with Slido' and the poll will launch automatically when you get to this slide.

BEFORE YOUR VOTE

Remember, the LoRA post it note is entirely private to the organisation

But also remember what organisational data includes. What do people suppose is in my inbox right now? Does the University own copyright in everything in my inbox or OneDrive?

RAG

The Return of the Copy: Unlike base training or fine-tuning, RAG does not decouple information from its structural building blocks. The source document survives entirely intact.

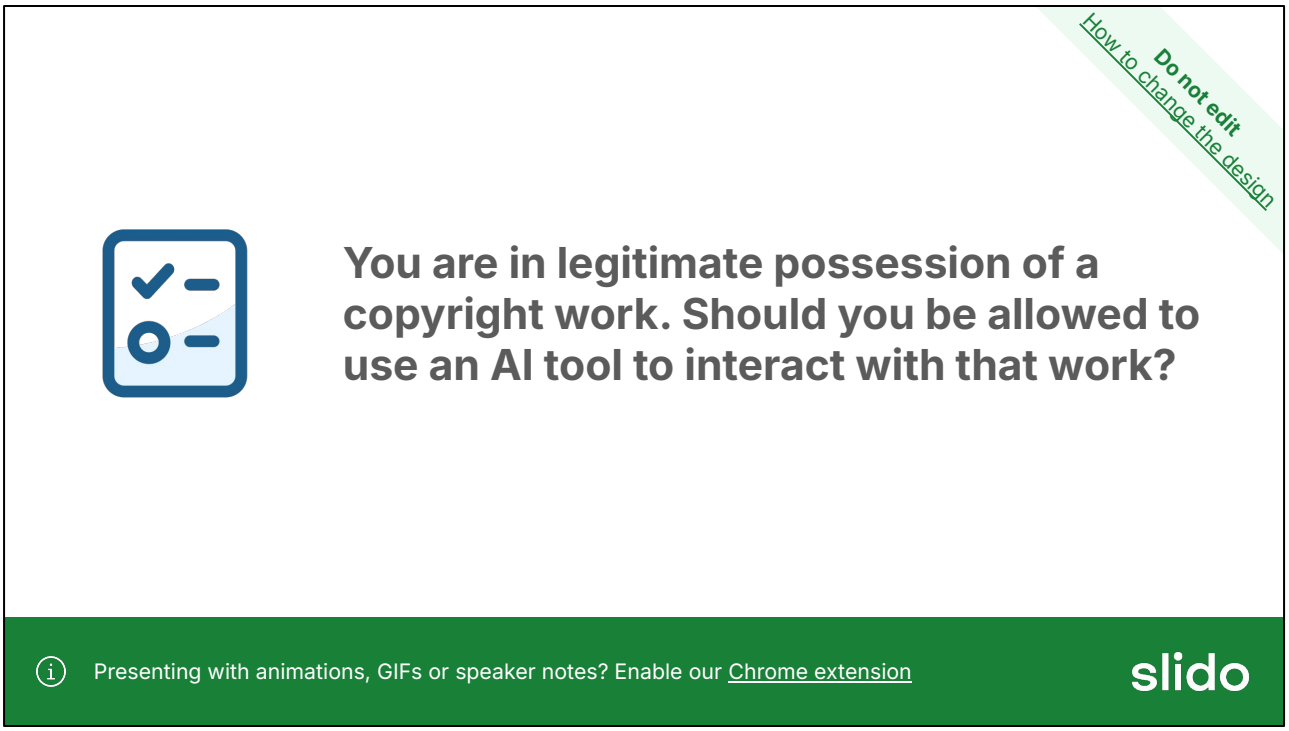
The Mechanism (The Digital Intern): There is no mathematical woodchipper here. RAG is a search-and-retrieve function. The system finds a relevant document in a database and pastes the text verbatim into the LLM's context window alongside the user prompt.

The Legal Reality: This is literal, unambiguous copying. It is traditional reproduction occurring dynamically in the context window. It skips the abstract copyright debates of "training" entirely.

The Industry "Open Secret" (SaaS Wrappers):

- Many "highly specialised" AI tools are simply generic base models wrapped in an automated RAG pipeline.
- They do not do bespoke training; they intercept your prompt, inject pre-canned instructions and curated documents into the hidden context window, and pass the result back to you.

The DIY Alternative: RAG is devastatingly effective, but it is not magic. With strategic prompting and your own reference documents, you can manually construct the exact same context window. You do not need to pay a subscription fee for a startup to execute a macro of Ctrl-C and Ctrl-V on your behalf.



📣 This is Slido interaction slide, please don't delete it.
✅ Click on 'Present with Slido' and the poll will launch automatically when you get to this slide.

Maybe the author has sent it to you. Maybe it is creative commons or crown copyright. Maybe the research and private study fair dealing exception applies.

Maybe you want a summary or critique? Maybe you want to translate it? Maybe you need to format shift to meet a physical or neurological need? Maybe it's been emailed to you or saved to a document management system that automatically does this to every document?

I know what the answer would be if the question was “Do you…”. I want to know whether you think this **should** be allowed.

Overfitting + memorisation

[Skipped for timing]

The Mechanism of Overfitting: If fine-tuning is briefly steeping a curated teabag to establish a general flavor profile, overfitting is leaving the teabag in the water until the model mathematically memorises the taste of the string, the staple, and the paper tag.

Statistical Failure vs. Storage:

- Overfitting is a failure of generalisation (vector collapse), not the retrieval of a saved file.
- The model has not "saved" the document; the training process has simply warped the local geometry so severely that the statistical probability of that specific, continuous sequence of tokens approaches 100%.

The Legal Reality (The Photocopier Problem): It largely does not matter to a judge how the text was generated. While technically a statistical anomaly rather than file retrieval, to a plaintiff's lawyer, verbatim regurgitation is functionally and legally indistinguishable from a photocopier.

Recommended reading:

- XKCD 2169 - Predictive models
- Berkeley AI Research (BAIR) 2019 - Quantifying Memorization in Neural Language Models

The Law

So what does the law have to say about all this!? Vast wealth has accrued to a few tech titans while authors and journalists are penurious. This will not stand! What does copyright have to say about this!? Not sure. I hope to learn all about that this time next week.

The Current Legal Landscape: Unresolved. The intersection of generative AI and traditional copyright is currently navigating a period of immense friction. Expect significant, protracted litigation over the next decade before definitive legislative updates emerge.

The Provenance Problem (How Models are Built):

- Base models rely on wholesale ingestion of the public internet.
- Significant friction exists regarding data acquisition mechanics—specifically the bypassing of paywalls and the ignoring of robots.txt exclusion protocols.

Dataset Contamination Risks:

- The industry faces ongoing scrutiny over the use of datasets with compromised or legally dubious origins (e.g., historical inclusion of sources like 'The Pile' or 'PiLiMi').
- Establishing clean data provenance is the central vulnerability in the current AI training paradigm.

The Golden Rule of Corporate Discovery: Internal developer communications acknowledging compromised data sources are discoverable in court. The most critical

lesson in legal risk management: do not document your own assumed liabilities in writing.

29 Infringement of copyright

(1) Copyright in a work is infringed by a person who, other than pursuant to a copyright licence, does any restricted act.

The Baseline Rule: Section 29 dictates that unauthorised copying is infringement. Conversely, possessing a license (permission) expressly permits the restricted act.

The Licensing Pivot: Major AI developers are increasingly executing multi-million dollar licensing agreements with massive data aggregators (e.g., Reddit, News Corp, Stack Overflow).

Strategic Implications (Regulatory Capture):

- While publicly framed as a pivot toward compliance or fair compensation, these exclusive licensing deals functionally establish a massive economic barrier to entry.
- If the legal consensus settles on "training requires licensed data," only entities with vast capital reserves can afford the toll to build foundational base models.

Market Consolidation: This dynamic severely disadvantages the open-source community and academic research departments, starving them of the "clean," legally unambiguous data necessary to compete.

The Incumbent Advantage: The pioneers of this technology built their foundational models on unconstrained public scraping. By now transitioning to and advocating for strict, highly expensive licensing regimes, they are effectively pulling the ladder up behind them.

But wait, there’s more

The Stateless Nature of LLMs: The mathematical model itself possesses zero memory. It does not dynamically train or update its core geometry in response to live user interactions. Once the context window is cleared, the session ceases to exist.

The Server Reality: While the model forgets, the corporate entity hosting the server absolutely does not. Every prompt, query, and uploaded document is logged and retained by the infrastructure provider.

The Contractual Trap: The transition from copyright law to contract law begins here. Most users never read the Terms of Service. This is a highly successful feature, not a bug, as tech providers deliberately obscure their data retention and usage policies deep within fine print.

Terms of Service and Data Use

…you (a) retain your ownership rights in Input and (b) own the Output…

We may use Content to provide, maintain, develop, and improve our Services…

When you use our [free] services, we may use your content to train our models. You can opt out of training through our privacy portal…

Chat GPT / Open AI

https://openai.com/en-GB/policies/row-terms-of-use/

https://openai.com/en-GB/policies/how-your-data-is-used-to-improve-model-performance/

Consumer Terms Privacy Policy

We may use Materials to provide, maintain, and improve the Services and to develop other products and services, including training our models, unless you opt out of training through your account settings.

Anthropic Claude

https://www.anthropic.com/news/updates-to-our-consumer-terms

https://www.anthropic.com/legal/consumer-terms

5 years: We are also extending data retention to five years, if you allow us to use your data for model training.

Change of tune: For a long time, a major Anthropic differentiator was a strict "we do not train on your data" stance. However, in late 2025, they updated their Consumer Terms and Privacy Policy, shifting how they handle data for their consumer tiers

Enterprise data protection statement

Your data is private: We won’t use your data except as you instruct.

Your data isn’t used to train foundation models: [Tool] uses the user’s context to create relevant responses. [Tool] also uses [RAG]... the prompts, responses, and data accessed through [RAG] aren't used to train foundation models.

Copilot

https://learn.microsoft.com/en-us/microsoft-365/copilot/enterprise-data-protection

Enterprise: Not entirely fair comparing an enterprise data protection commitment and “free” tools. But this commitment applies to Copilot Chat, the free version available at the University if you log in with your staff or student account.

Really Free: Genuinely free in the sense of neither dollar nor data cost. However corporate IT spends a lot of time and effort to manage.

Privacy and Terms of Use

The content in [tool name] will not be used to directly train our foundational AI models, unless you choose to provide feedback.

Google Notebook LM

https://support.google.com/notebooklm/answer/17004255

https://support.google.com/notebooklm/answer/16164461

The "Deep Work" Paradigm: Divergence in general vs “deep work” environments. For deep work / enterprise, absolute data privacy is non-negotiable for enterprise data.
The RAG Reality: RAG architecture. May not need to train the base model on your data?

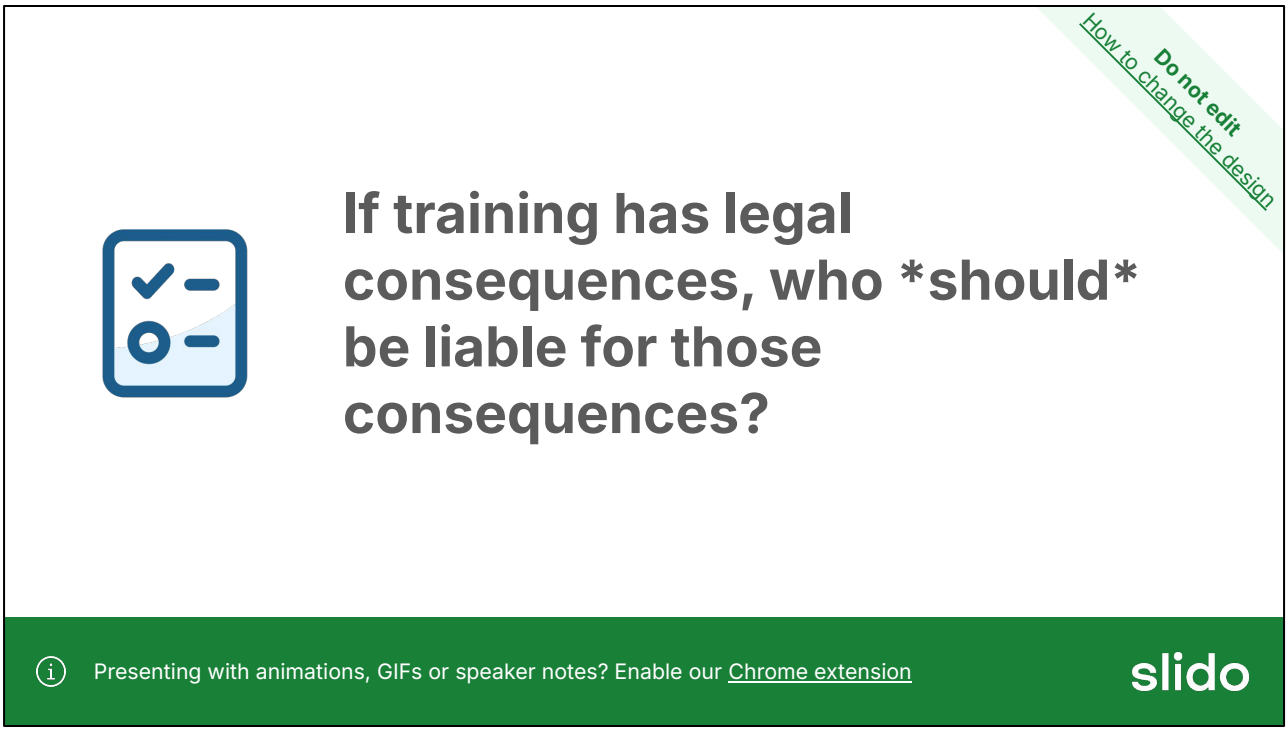
Common across all: Contract trumps copyright

The Ultimate Bypass: Bypass the copyright debate using pure contract law.

The Two-Way Contractual Shield: The Terms of Service operate as an absolute override of statutory copyright:

- **Training is Copying:** If the courts eventually rule that training requires permission, the ToS already contains a clause where you clicked "I Agree" to license your data to them.
- **Training isn’t Copying:** Even if the courts rule that training is entirely legal, the AI company use contract law to promise not to train on your data. Provided you pay a premium for their "private" enterprise tier.

Conclusion: The actual rules governing generative AI are no longer found in copyright legislation. They are dictated entirely by the private agreements we blindly click to access.



📣 This is Slido interaction slide, please don't delete it.
✅ Click on 'Present with Slido' and the poll will launch automatically when you get to this slide.

Slightly different rating scale.

- **AI companies definitely at fault?** If they really mean this in their terms and conditions, they need to do more to educate and moderate what is uploaded. There are guard rails to filter if they are producing copyright text, technically they can filter at least some uploads too.
- **User?** You probably kind of know you shouldn't be doing some of this stuff. Yeah, they could be clearer. But you clicked agree.

It gets worse

Read your terms of service!

Terms of use

You are responsible for Content, including ensuring that it does not violate any applicable law or these Terms. You represent and warrant that you have all rights, licenses, and permissions needed to provide Input to our Services.

… you will indemnify and hold harmless us, our affiliates, and our personnel, from and against any costs, losses, liabilities, and expenses (including attorneys' fees) from third party claims arising out of or relating to your use of the Services and Content or any violation of these Terms.

The Ultimate Risk Shift: We established that the Terms of Service bypass statutory copyright. This clause is the enforcement mechanism. It transfers the entirety of the legal risk from the platform provider to the end-user.

Warranties: "I am so certain of this fact that I accept very strict liability if it turns out to be untrue." By clicking "Agree," the user legally represents and warrants that they possess the absolute right, license, and permission to provide the inputted data.

The Indemnity (The Blank Check): An indemnity is not merely a promise to behave; it is a strict financial liability mechanism. It dictates: "If my data input causes this tech titan to be sued by a third party, I agree to personally pay their legal fees, their court costs, and any resulting settlement damages."

The Hidden Catch (The Licensing Trap): It is not enough to possess the right to *read* or *use* a document. By uploading it, you are warranting that you have the right to authorise the AI company to ingest and process it under their terms. If you upload third-party IP without that specific authorisation, the indemnity is triggered.

Business only. Does not apply to individuals (may not be enforceable). But "organisation" includes Universities.

More law

Non disclosure agreements

Privacy Act

Export controls

Te Herenga Waka Victoria University of Wellington Generative AI Policy

NDAs & Confidentiality (The Mathematical Disclosure):

- Does converting a confidential document into floating-point weights constitute a legal disclosure?
- Even if the base model claims it is "unretrievable," the risk of probabilistic reassembly (overfitting/memorisation) creates a massive, unquantifiable breach risk.

The Privacy Act:

- We have strict obligations to protect Personally Identifiable Information (PII).
- If PII is mathematically woodchippered during training, is it legally "anonymised"? You cannot base statutory privacy compliance on the "odds" that the model won't regurgitate it.

Export Controls:

- Export laws strictly control the transfer of dual-use technologies (e.g., military, drone research).
- Feeding sensitive engineering blueprints into a public, offshore LLM API is functionally equivalent to exporting it.

VUW Policy:

- The University's Gen AI Policy dictates acceptable use and data classification - safe AI tools only for sensitive, restricted, private and third party IP information.

Many other potential laws: Financial regulation (e.g. insider trading), the Official Information Act (OIA), Public Records, Evidence Act.

“Are you really going to litigate whether 'probabilistic geometric reassembly' constitutes a breach of an NDA? Good luck billing the client for that”

- Gemini

More slides!? This can’t be good.

General Terms

…you must not (i) run or install any computer software or hardware on, against, in relation to, or as an overlay over, our Services or network; (ii) mine, scrape, index, or otherwise automatically access, collect, copy, download or record the Property; or (iii) automatically connect (whether through APIs or otherwise) the Property to other data, software, services or networks…

Thompson Reuters General Terms clause 3(f)

https://www.thomsonreuters.com/content/dam/ewp-m/documents/thomsonreuters/en/pdf/other/general-terms-nz-q225.pdf

The Equal Opportunity Audit: We've roasted the AI companies for weaponising contracts. Now let's look at the traditional data gatekeepers using law as an example.

General Terms and Conditions

… you are… prohibited from downloading, storing, reproducing, transmitting, displaying, printing, copying, distributing, or using Materials retrieved from the Services. You may not print or download Materials without using the printing or downloading commands of the Services or your web browser software. All access to and use of the Services via mechanical, programmatic, robotic, scripted or any other automated means not provided as part of the Services is strictly prohibited. Use of the Services is permitted only via manually conducted, discrete, individual search and retrieval activities.

Lexis Nexis General Terms and Conditions clause 1.3

https://www.lexisnexis.com/terms/lngp/nz/

The Anti-Automation / Robot Paranoia: Target all AI access methodologies:

- **The Anti-Agent Provision (No Overlays):** Express prohibitions against running software "over" their interface. This is a targeted ban on client-side AI: no agentic web browsers, no screen viewers, no accessibility, extensions reading the screen alongside you. No smart glasses?
- **The Anti-Ingestion Provision (No Scraping):** Absolute prohibitions on corpus extraction and mass downloading to starve base model training.
- **The Anti-RAG Provision (No APIs):** Banning automated connections to their databases, ensuring your digital intern cannot use Ctrl-C.

Humans Only: Beyond specific software, they explicitly demand that retrieval be "manually conducted" and "discrete." You are contractually forbidden from extracting knowledge using AI.

Copyright: But New Zealand primary law (statutes, case law) CC0 / public domain / not copyrighted. Because legacy publishers cannot use copyright to prevent AI from reading public laws, they use contract law to build a tollbooth around the access point.

The Walled Garden Upsell? Enforcing a strict monopoly over how you interact with

public knowledge, setting stage for own tools.

Should customers be allowed to use AI tools on third party sourced data they have the legitimate right to use?

📣 This is Slido interaction slide, please don't delete it.
✅ Click on 'Present with Slido' and the poll will launch automatically when you get to this slide.

“They are using pure contract law to ensure the only intelligence allowed to process their premium data is the easily-billable, carbon-based kind.”

- Gemini

Show of hands

Where should we be drawing the line between allowed and banned behaviour for end users?

- **Email and document management with smart features**
- **Agentic web browsers**
- **Private LLMs**
- **Public LLMs, if training is off**
- **Public LLMs, contractual assurances that training is secure**
- **Public LLMs, all training**

Where is the line?

Who just wants to ban lawyers and legal terms and conditions instead?

The Spectrum of Acceptable Use: Where do you draw the line on risk?

- *Low Risk / Walled Gardens:* Email/document management with smart features.
- *Medium Risk / Private Enclaves:* Private LLMs and Public LLMs with strictly enforced "no training" contracts.
- *High Risk / The Wild West:* Agentic web browsers and Public LLMs with training enabled.

Thesis: Generative AI training is high-dimensional trigonometry. It is a destructive, one-way statistical process that fundamentally defies traditional legal concepts of "copying." But the debate over statutory copyright definitions may be becoming increasingly irrelevant. While the courts spend the next decade debating whether maths is illegal, the industry has already attempted to bypass the problem using Terms of Service. By the time definitive legislative answers arrive, we may be locked into proprietary, contractually-fenced AI agents, ironically provided by the very entities who built their foundational models on unconstrained scraping.

The Handoff: If you're really interested in actual law... come back next week. I'm sure Graeme and Daniel have some much more legal opinions on the matter than I do.

“Debating the nuances of statutory copyright in 2026 is basically just LARPing for lawyers who don't understand how APIs work.”

- Gemini

I’ll leave the last word to my coauthor.