SREDStakeholder.CA

Best AI Tools for Scientific Research & SR&ED

Research quality • Productivity • Confidentiality • Practical tool selection

A function-first guide for Canadian researchers and SR&ED claimants

Research quality
Evidence & verification

Productivity
Structured workflows

Confidentiality
Secure deployment

Selection
Task-specific toolkits

Authored & presented by David Sabina, MEUK Corporation

Nov. 26, 2026 • SREDStakeholder.ca webcast

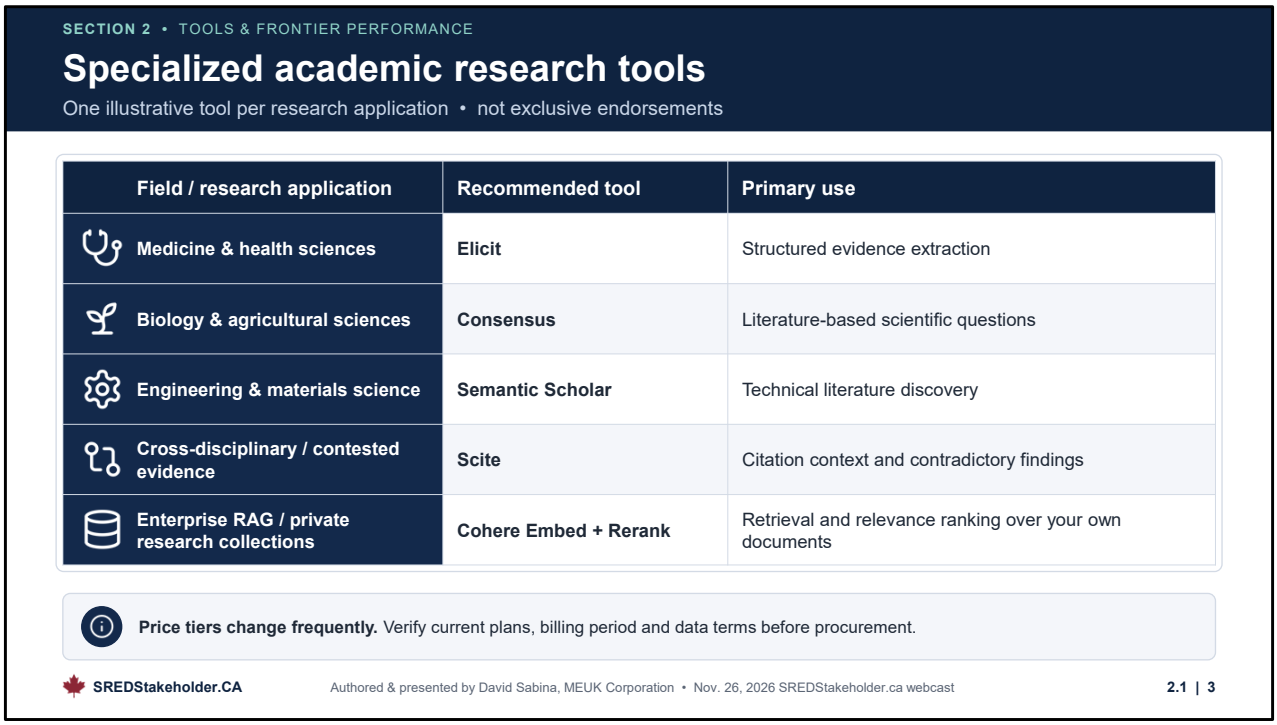
1

Presentation contents

Five sections • scientific research and SR&ED

1	**Research framework**	Research quality, human–AI collaboration, and academic evidence
2	**Tools & frontier performance**	Academic overlays, Cohere, model families and specialized benchmarks
3	**Maximizing research performance**	Prompting, extraction, validation and multi-model workflows
4	**Confidentiality & deployment**	Canadian providers, deployment tiers and inference harnesses
5	**Customization & SR&ED**	Model weights, harness modifications, qualifying uncertainty and takeaways

The presentation groups tools by research function, evaluation quality, deployment confidentiality, and potential SR&ED. Benchmarks are directional, not substitutes for project-specific tests.



One illustrative recommendation is provided for each of five research applications; these are not exclusive endorsements. Elicit supports structured extraction from scholarly studies (elicit.com). Consensus supports literature-grounded scientific questions (consensus.app). Semantic Scholar offers scholarly search and citation metadata (semanticscholar.org). Scite adds citation-context analysis (scite.ai). Cohere Embed and Rerank are enterprise retrieval and reranking models rather than independent academic literature databases: users must supply or connect an authorized document corpus (cohere.com). Tool capabilities, pricing and data-handling terms should be checked for the selected deployment.

SECTION 2 • TOOLS & FRONTIER PERFORMANCE

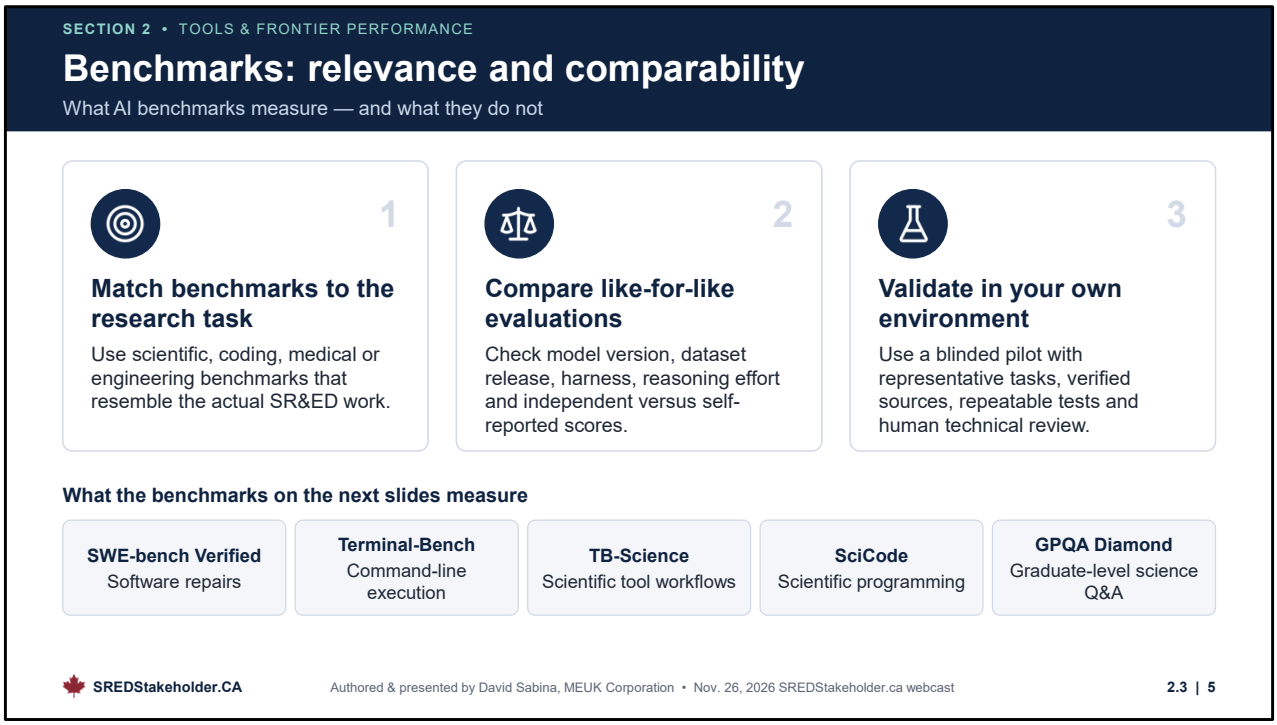
Why combine academic overlays with frontier models?

What a specialized research index adds to a general-purpose model

Function	Frontier model alone	Overlay + frontier model
Discover evidence	Flexible web and scholarly searches	Specialized indexes and citation networks
Evaluate literature	Summarize selected papers	Screen studies; extract methods and outcomes
Check competing claims	Prompt for contrary findings	Citation context and systematic comparisons
Develop hypotheses	Reason, model, code and plan	Reason over a curated, traceable evidence base
Document SR&ED	Record reasoning and actual tests	Retain search strategy, sources and decisions

Use an overlay when it adds coverage, extraction, citation context or reproducibility; direct model retrieval may suffice for simple questions.

Academic search services can add scholarly indexing, citation networks, extraction fields, and reproducible review workflows. Frontier models may also retrieve papers directly; the additional value of an overlay depends on coverage, source fidelity, extraction quality and cost. Neither approach guarantees that a citation supports the claim made. Sources: https://www.semanticscholar.org/ ; https://elicit.com/ ; https://scite.ai/



Benchmark results measure distinct tasks in different research domains. SWE-bench Verified concerns repository-level software repairs. Terminal-Bench concerns command-line tools and multi-step execution. Terminal-Bench-Science targets scientific tool workflows. SciCode concerns numerical and scientific programming. GPQA Diamond measures graduate-level questions in physics, chemistry and biology. Results depend on benchmark version, model configuration and evaluation setup. These tests do not establish literature-retrieval fidelity or the validity of actual experiments. Cohere Command A+ figures on the following slides are supplied estimates, because directly comparable verified results were not established for this specific benchmark set. Sources: https://www.swebench.com/ ; https://www.tbench.ai/ ; https://scicode-bench.github.io/ ; https://livebench.ai/ ; https://docs.cohere.com/docs/command-a-plus

SECTION 2 • TOOLS & FRONTIER PERFORMANCE

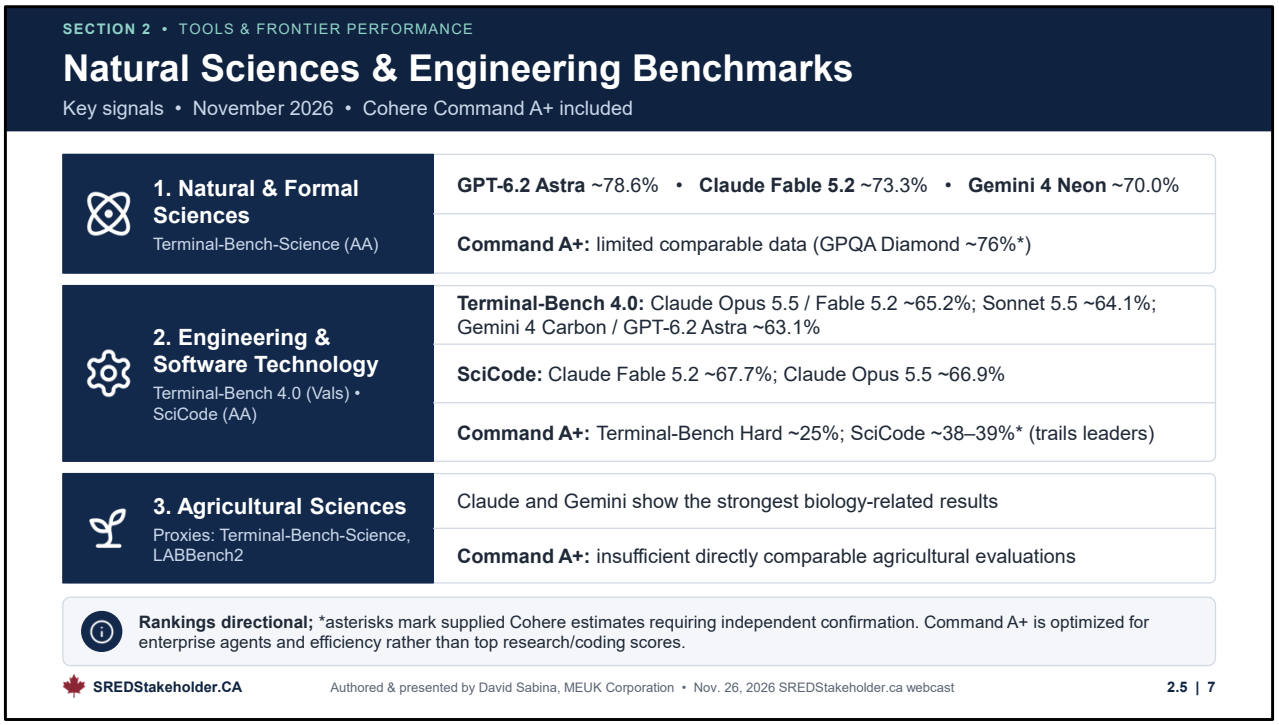
Frontier AI: Benchmark Performance Comparison

Reported approximate scores • November 22, 2026 • different versions, harnesses and test setups

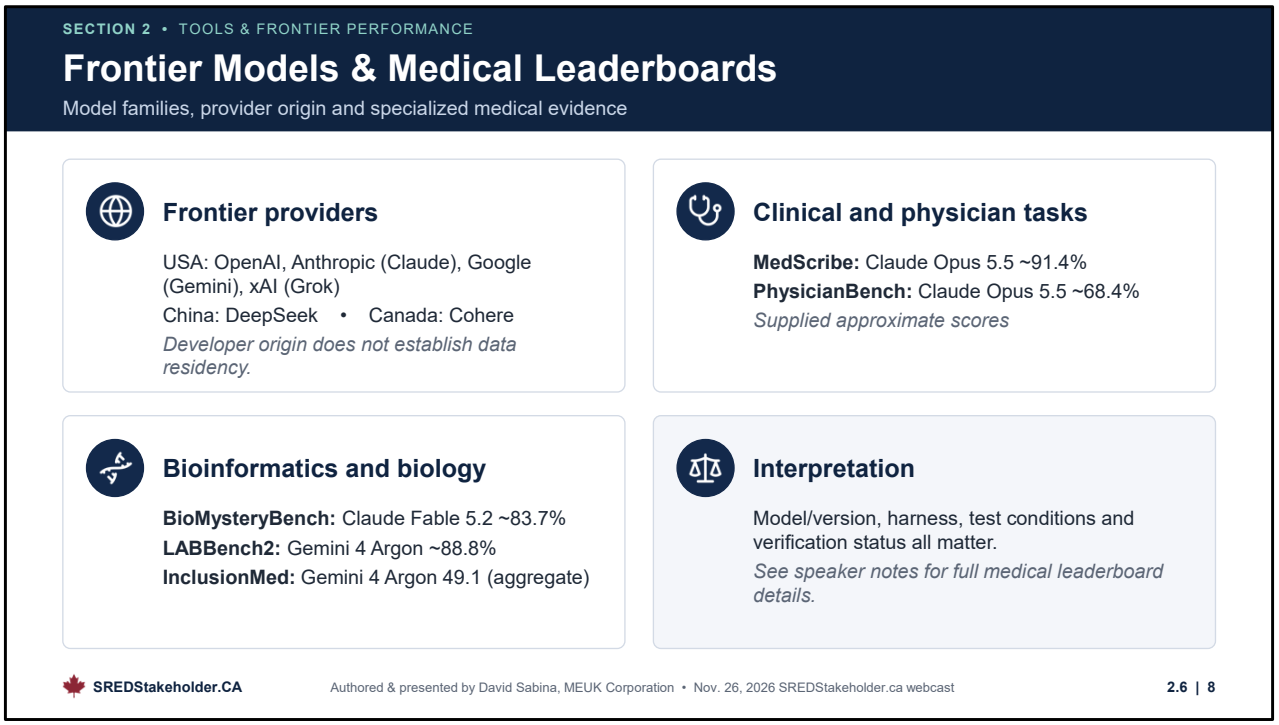
Benchmark	Grok (4.6 / 4.7)	Claude (Fable / Opus)	OpenAI (GPT-6.x)	Gemini (4 series)	DeepSeek (V4 Pro)	Cohere (Command A+)
SWE-bench Verified	~95.6%	~95–96%	~96%	~75–81%	~96.4%*	Limited data
Terminal-Bench 4.x	~20–38%	~65%	~63%	~63%	~15–32%	~1–25%*
Terminal-Bench-Science	~7–14%	~73%	~78%	~70%	~10–16%	Limited data
SciCode	~57%	~63–67%	~56–58%	~62%	~50–55%	~38–39%*
GPQA Diamond	~93%	~93–96%	~96%	~94–97%	~92%	~76%*
Provider country	USA	USA	USA	USA	China	Canada

- Scores are approximate and vary by agent harness, reasoning effort and evaluation setup — do not compare across harnesses.
- *Cohere and DeepSeek figures shown are supplied estimates, not independently verified.
- Cohere emphasizes enterprise agents and efficient private deployment; missing scores are not zeros.
- Provider country is not the same as data-hosting location.

Benchmark figures are approximate compilations supplied for November 22, 2026. SWE-bench Verified evaluates software issue resolution; Terminal-Bench evaluates terminal-based agent execution; Terminal-Bench-Science addresses scientific workflows; SciCode examines scientific programming; GPQA Diamond measures challenging graduate-level scientific knowledge. Results depend on harness, reasoning effort, model version and independent versus self-reported evaluation. Cohere Command A+ has a different enterprise-oriented performance profile; unverified scores should not be interpreted as independently established. Reference sites include https://www.swebench.com/ and https://www.tbench.ai/ .



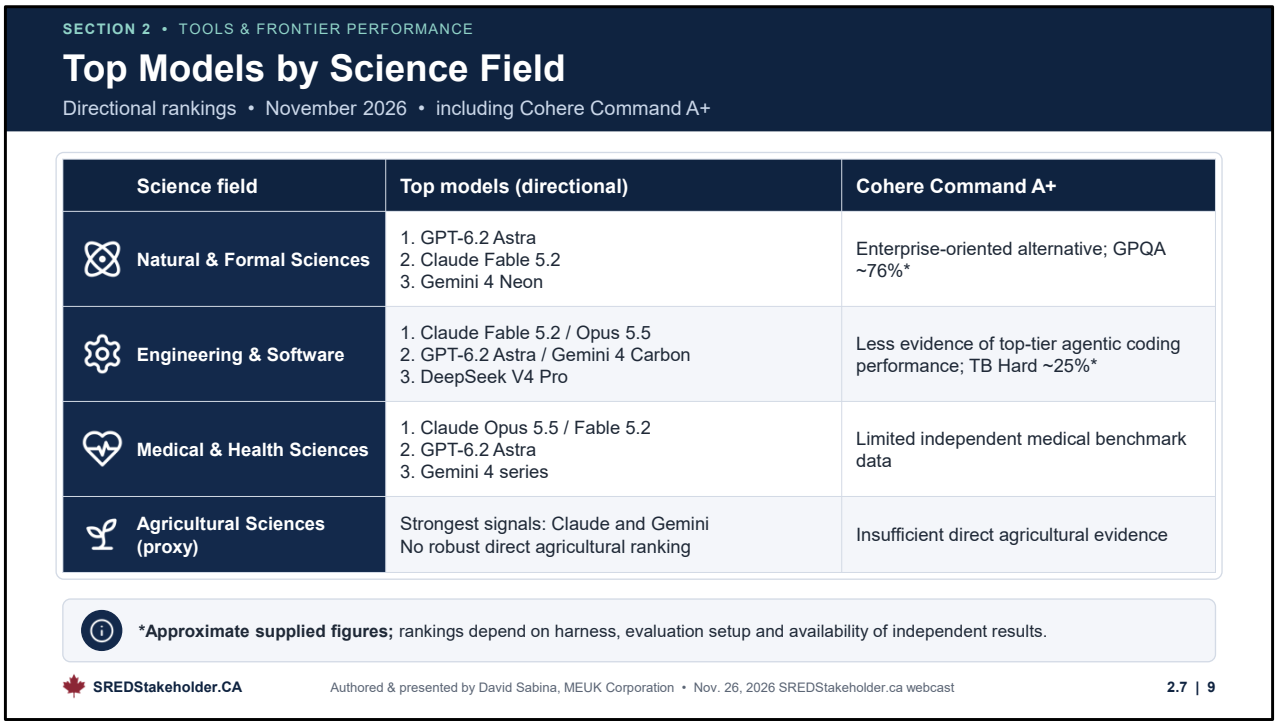
Terminal-Bench-Science relates to scientific terminal workflows, while Terminal-Bench 4.0 evaluates agentic execution in technical environments. SciCode tests scientific coding and is grouped under engineering and software technology. Agricultural comparisons use life-science proxy benchmarks rather than direct agricultural task evaluations. Cohere Command A+ should be compared for enterprise retrieval, agent and deployment requirements as well as raw benchmark results.



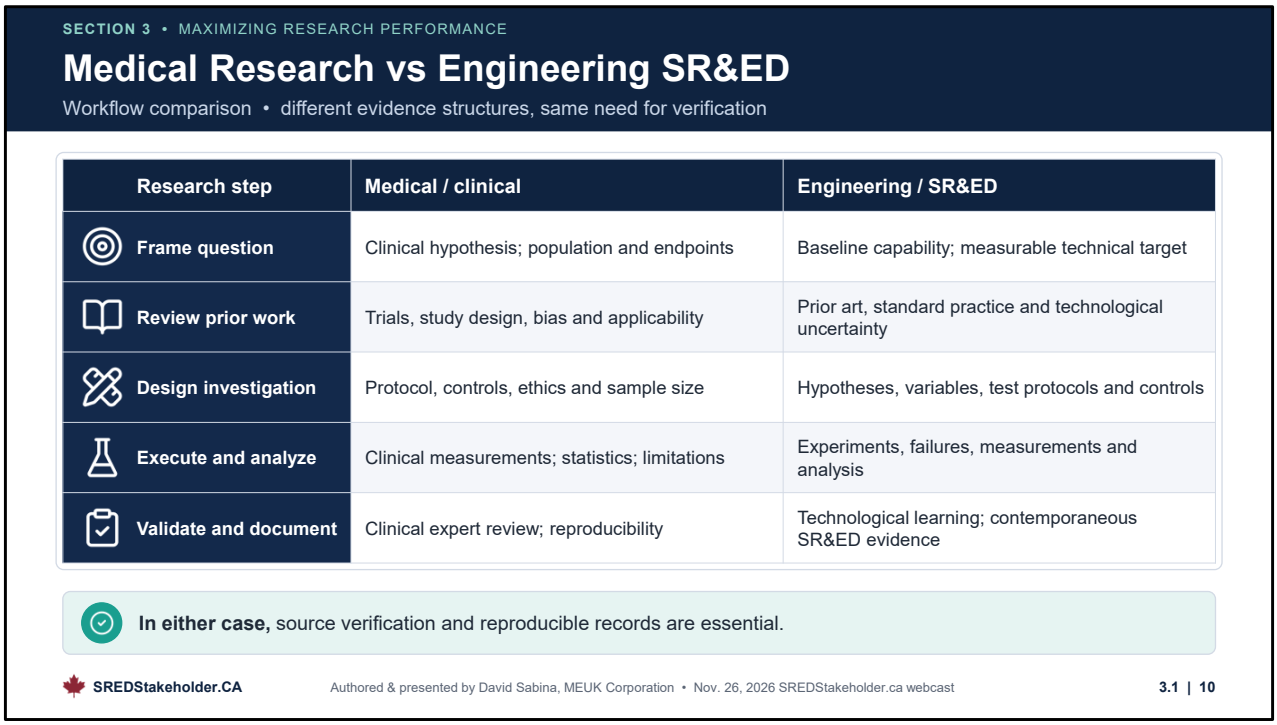
Frontier-model comparison as of November 22, 2026. The provider country is the location of the developer, not a statement of the location of inference, storage or training. Cloud processing locations depend on region, platform, routing, contract and enterprise configuration. Cohere is headquartered in Canada. Command A+ (released May 20, 2026) is a 218-billion-total / 25-billion-active parameter mixture-of-experts model with a 128K context window, 48 supported languages, vision input, tool use and reasoning. Its model weights are offered under Apache 2.0, and Cohere documents private deployment options. The other model-family names are retained from the proposed preview and should be verified against each developer's current official release notes before public presentation. Sources: https://docs.cohere.com/docs/command-a-plus ; https://docs.cohere.com/v2/changelog/command-a-plus-05-2026 ; https://cohere.com/models-overview .

Medical leaderboard detail:

MedScribe concerns clinical note generation; PhysicianBench concerns electronic health record and physician workflows; BioMysteryBench tests agentic bioinformatics investigations; LABBench2 concerns biology research with literature, databases and bioinformatics; InclusionMed is an aggregate medical portfolio. Scores shown are supplied approximations, not a single independently normalized evaluation. Clinical performance cannot establish medical safety or suitability without validation and professional oversight.



The model rankings are directional and drawn from different evaluation harnesses and sources. Natural/formal sciences, engineering/software, medical/health, and agricultural sciences differ in their task distributions. The Cohere column reflects the supplied approximate benchmark context, not a complete independent ranking. Field-specific pilots and source verification remain necessary.



Clinical research and engineering SR&ED require different evidence structures. Clinical work commonly emphasizes study design, endpoints, confounding, ethics and interpretation; engineering work emphasizes technological uncertainty, experimental variables, measured performance and technological advancement. In either case, source verification and reproducible records are essential. Source: CRA SR&ED program guidance: https://www.canada.ca/en/revenue-agency/services/scientific-research-experimental-development-tax-incentive-program.html

SECTION 4 • CONFIDENTIALITY & DEPLOYMENT

Confidentiality: Deployment Tiers for AI Models

Match the deployment tier to the sensitivity of the research data

Tier	Typical controls	Example providers / deployment	Residual risk	Most relevant to
Tier 1 **Consumer**	Settings + data minimization	ChatGPT; Claude; Gemini (consumer plans)	Vendor-policy exposure	Open research; avoid proprietary or regulated data
Tier 2 **Enterprise / API**	Contract, retention limits, access controls	OpenAI API / Enterprise; Anthropic Claude Enterprise; Google Vertex AI	Configuration and processor risk	Organizations needing stronger assurance
Tier 3 **Private hosting**	Approved Canadian / private infrastructure	Cohere private deployment; Azure-hosted models; self-hosted open models	Operations, backups, access controls	Institutions requiring stronger residency or control
Tier 4 **Offline**	Local models + restricted networking	Ollama; LM Studio; Open WebUI with locally deployed models	Device, logs and physical access	Highest-sensitivity research; local security burden

Illustrative tiers, not formal security ratings. Verify contracts, logging, subprocessors, region and institutional approvals.
Private hosting ≠ automatic Canadian residency. Offline systems still carry endpoint, supply-chain and physical risks. Source: cyber.gc.ca

Deployment tiers describe different technical and contractual control arrangements, not security ratings. Consumer tools generally require strong data minimization. Enterprise services may offer contractual retention limits, access controls and audit features. Private hosting can improve operational control, but Canadian vendor headquarters alone do not guarantee Canadian data residency. Cohere deployment and regional availability must be checked for the specific product and agreement. Offline deployments reduce external data transmission but shift patching, access control, monitoring and hardware security responsibilities to the organization. Source: Canadian Centre for Cyber Security, https://www.cyber.gc.ca/ ; vendor service and contractual documentation.

SECTION 4 • CONFIDENTIALITY & DEPLOYMENT

AI Inference Harnesses: Functions and Trade-offs

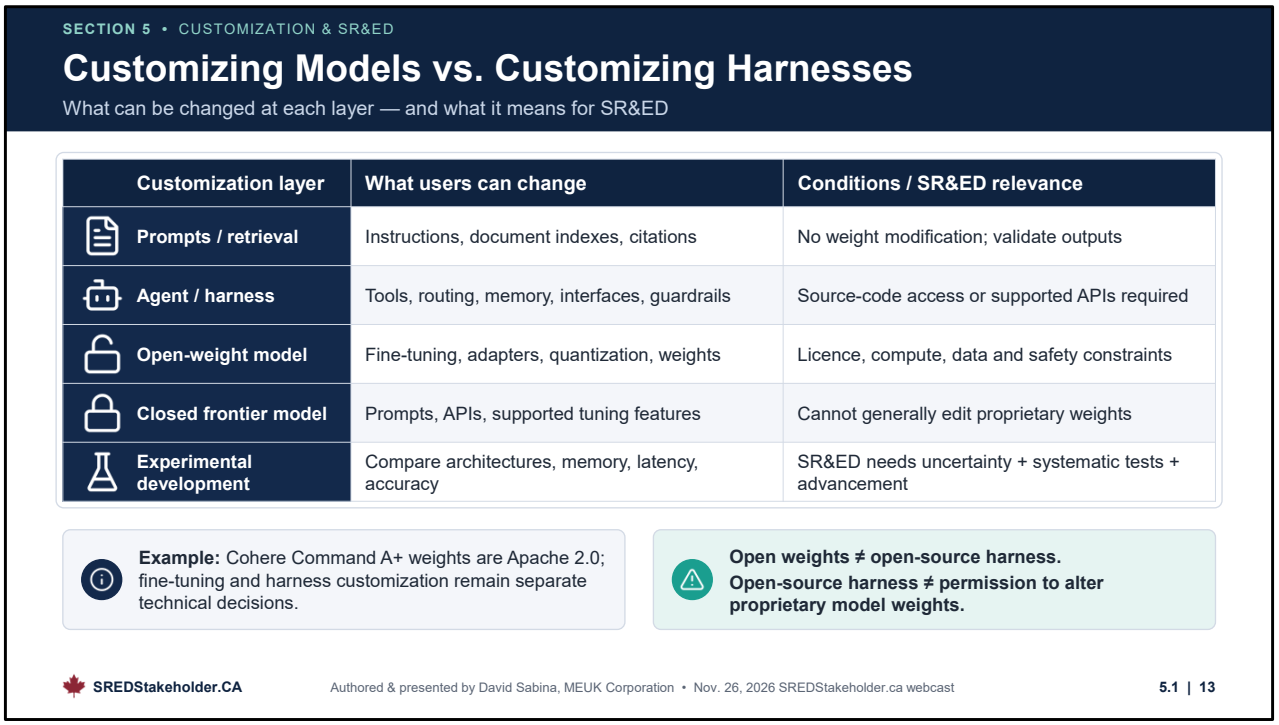
Local and low-memory options for running models

Harness = software that loads a model, manages memory and inference, exposes interfaces, and may orchestrate tools or agents.

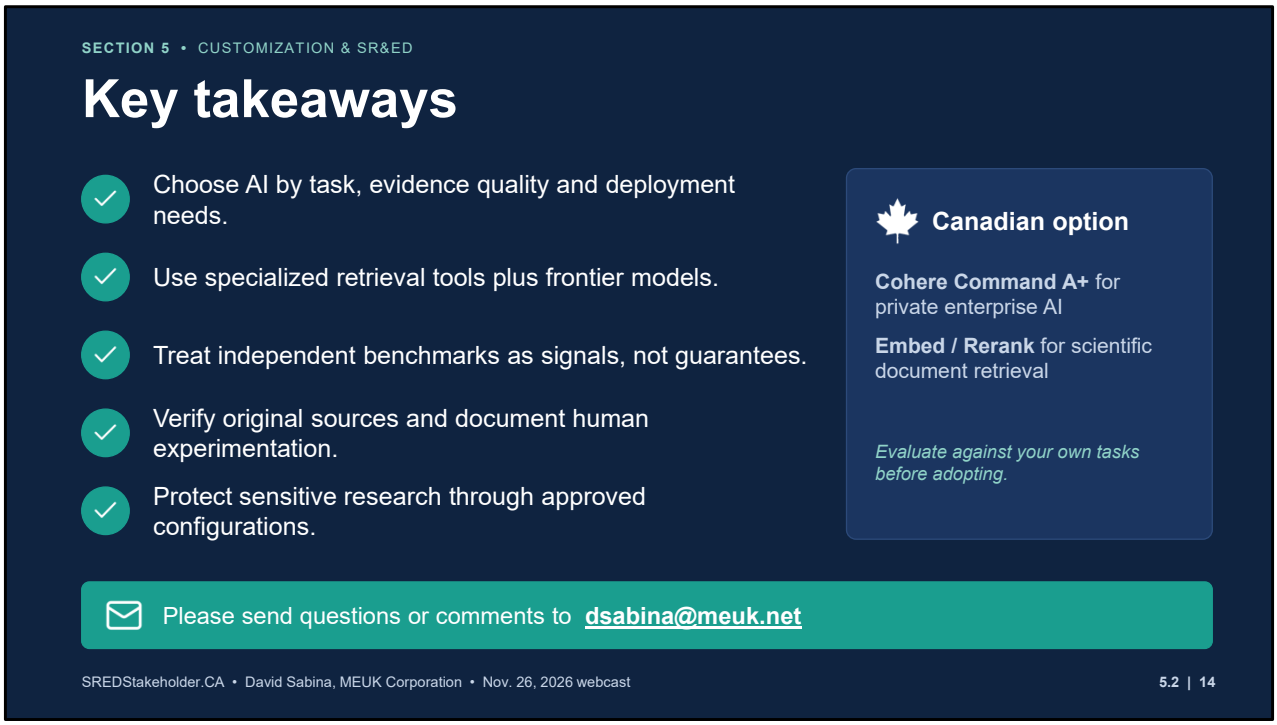
Harness / layer	Why choose it	Trade-off / caution
Ollama	Easy local model serving and API	Compatibility / performance vary
LM Studio	Desktop model testing and GUI	Less suited to complex production orchestration
Open WebUI	Customizable chat / retrieval interface	Not itself a low-memory inference engine
AirLLM	Layer / expert streaming for limited VRAM	Storage I/O and latency; model support varies
Colibri+ / AirColibri*	Experimental tiered / streamed low-memory execution	Immature support; verify project and benchmarks

***Naming and implementations vary; AirColibri is explicitly experimental.** Low VRAM does not mean low total compute or fast execution.

An inference harness is the software surrounding a model: loading model files, allocating memory, tokenization, inference scheduling, caching, quantization support, API serving and sometimes tool or agent orchestration. Ollama emphasizes simple local serving; LM Studio emphasizes desktop model exploration; Open WebUI is a user interface and orchestration layer rather than a memory-saving inference engine. AirLLM streams layers to reduce GPU-memory needs, potentially increasing storage I/O and latency. Colibri-related implementations require project-specific verification of supported hardware, licenses and performance. Harness code can be modified when source and license permit, or extended through APIs. Model weights are independently governed by their own licenses. Cohere Command A+ weights are Apache 2.0, but local deployment has significant GPU requirements (Cohere documents 1 B200 or 2 H100 at W4A4); low-memory harness claims must not be assumed to make this model run efficiently on ordinary laptops. Sources: https://docs.cohere.com/docs/command-a-plus ; https://ollama.com/ ; https://lmstudio.ai/ ; https://openwebui.com/ ; https://github.com/IRIADA/airllm .



Model weights are learned numerical parameters. Open-weight availability may permit local execution and, depending on license, fine-tuning, quantization or modification. A harness is software around the model: an open-source harness can often be changed independently of the weights. A proprietary hosted model may permit prompt engineering, retrieval, function calling and supported fine-tuning without allowing direct weight access. Conversely, a model's open weights do not guarantee the associated runtime or service is open source. SR&ED eligibility is not established by customization alone; technological uncertainty, systematic investigation and technological advancement must be supported by evidence. Source: https://www.canada.ca/en/revenue-agency/services/scientific-research-experimental-development-tax-incentive-program.html



The principal selection factors are evidence quality, task fit, model access and deployment controls. Cohere offers an additional Canadian option: Command A+ for enterprise reasoning and agents, and Embed 5 / Rerank 5 for document retrieval. Its general-purpose benchmark standing should be tested rather than assumed. Sources: https://docs.cohere.com/docs/command-a-plus ; https://docs.cohere.com/docs/rerank .