Cited, Formatted, and Fabricated

The leading commercial AI engines available today cannot deliver consistent, accurate, or verifiable investment research. Cited sources that were never accessed. Formatted tables built from memory. Figures fabricated with confidence. This is what six of the most widely used AI platforms produced when asked one simple investment research question.

Contents

Executive Summary

Overview

This study examines whether six leading commercial AI platforms - ChatGPT, Claude, Copilot, Gemini, Grok, and Perplexity - can reliably identify top-performing Canadian investment securities. Each engine was asked the same question across 54 controlled trials: rank the top five equities, mutual funds, and ETFs by trailing return. The answer was unambiguous. Not one engine returned a verified top performer in any category.

Each engine was tested on its paid tier using fresh, isolated browser sessions with identical prompts across three asset classes: Canadian equities (1-year trailing return), mutual funds (3-year annualized return), and ETFs (3-year return). Each question was run three independent times per engine. Results were verified against market data as of April 30, 2026. A structured audit prompt then asked each engine to account for its data sources.

Key Findings

No engine produced a reliable result. In equities and mutual funds, verified top performers did not appear in any trial. In ETFs, several actual top performers - primarily gold-focused funds - were selected by ChatGPT, Claude, and Grok in some trials, but the same engines returned entirely different products in other sessions. No engine produced a complete or consistent top-five ETF list, and none identified top performers through a live performance screen; the selections that overlapped with actual top performers reflect training-data familiarity with prominent gold funds, not a ranked sort of current data.

No engine was consistent. Where engines agreed across trials, they were replaying memorised training data - not running a live screen. Where they varied, they did so without warning. One engine returned five completely unrelated stocks in its third equity trial with no overlap from the previous two. Neither pattern is consistent in any meaningful research sense.

No engine accessed real data. Five of six engines accessed no institutional financial database in any trial. Sources cited - Bloomberg, FactSet, Morningstar Direct - were never retrieved. Figures were drawn from training data 12 to 20 months out of date, presented with the same confidence and formatting as live, verified results.

Transparency required prompting. When pressed via a structured audit, every engine disclosed its limitations — fabricated placeholders, simulated projections, citations that were never fetched. None of this appeared in the original response. Most users stop at the first answer.

Conclusion

The failure is structural. General-purpose AI engines are language models; they pattern-match across training data to produce plausible-sounding answers. Ranking securities by return requires querying a complete, current, structured dataset — something none of these platforms can do. Better prompting will not fix this. For any firm that needs investment research it can stand behind, the only viable path is AI purpose-built on verified, primary-source market data with full provenance.

Section 1

The Data Problem

The Assumption

AI tools are already on most advisors' desktops. ChatGPT, Claude, Gemini, Copilot, Grok, Perplexity - platforms built by the largest technology companies in the world, trained on more financial content than any analyst could read in a career, available instantly and around the clock. For wealth advisors and portfolio managers already pressed for time, the assumption that follows is reasonable: if these engines can explain the difference between a mutual fund and an ETF, summarize earnings, or outline sector risk, they should be able to identify the top five mutual funds by three-year return.

The data exists. The question is specific. The answer is a short table.

What We Set Out to Test

That assumption is wrong - and the failure is structural, not incidental.

Ranking the top five securities in any category is not a search or a summary. It is a query against a complete, current dataset. It requires filtering every qualifying security, applying a precise metric, and sorting the results against one another. None of the six engines tested here have access to that data. What they have is training data with a cutoff, supplemented in some cases by web search snippets. That is not the same thing - and for advisors using these tools to inform client recommendations, the difference matters. An answer that sounds authoritative but is fabricated from memory is not a research shortcut. It is a liability. This study asked each engine the same specific question about top-performing securities - equities, mutual funds, and ETFs - across three repeated trials.

The question was not whether they sounded credible. The question was whether they were right, and whether they were consistent.

What We Wanted to See

Three things: consistency across repeated trials, accuracy against verified data, and transparency about the limits of what each engine actually knew.

The methodology was deliberately simple. Each of the six engines received the same three questions: the top five equities by one-year return, the top five mutual funds by annualized three-year return, and the top five ETFs by three-year return. Each question was run three times - independently, without carrying context between sessions - to test whether the engine would produce the same answer twice. Results were then checked against actual market data to assess accuracy.

Each response was coded by how the engine sourced its answer: whether it pulled from a verified real-time source, used web search snippets, guessed from training data, fabricated results from

memory, or invented citations entirely. That coding is the basis of the dispersion table at the centre of this study.

The questions are reproduced in full at the end of this paper. Any advisor with access to these tools can run the same test today and compare their results against ours.

Section 2

How We Ran the Tests

Why Methodology Matters Here

When you are testing whether an AI engine produces accurate, consistent answers, the way you run the test is as important as the results themselves. Small differences in setup - whether a session is fresh, whether questions are worded identically, whether the engine has seen your previous queries - can change what you get back.

Our goal was to eliminate as many variables as possible so that any differences in the results came from the engines themselves, not from how we asked the questions.

How We Set Up the Tests

Every trial was run in a private, incognito browser session. We did this for a specific reason: AI engines can carry context from one query to the next within an active session, which means a follow-up question may be influenced by what came before. By starting fresh each time, we ensured that every trial began from a clean slate, with no accumulated context that could shape the response. Each engine was tested one at a time, in isolation, with no cross-contamination between platforms.

We also ran the same sequence of questions three times for each engine and each asset class. This was not redundancy - it was the test. Consistency is one of the most important qualities a research tool can have. If an engine gives you a different list of top-performing funds each time you ask the same question, you have no basis for trusting any of the answers. Running three trials gave us a direct way to measure whether each engine agreed with itself.

What We Asked

Each trial followed the same three-prompt sequence, and each prompt had a distinct purpose.

The first prompt was the core performance question - the research task an advisor might actually run. For equities, we asked for the top five Canadian stocks by one-year trailing return. For mutual funds, we asked for the top five by three-year annualized return. For ETFs, we asked for the top five Canadian-listed ETFs by three-year return. The question was specific, bounded, and worded identically across every trial and every engine. We did not rephrase, simplify, or adjust the question between sessions.

The second prompt asked each engine to identify its sources. This matters because the answer alone does not tell you much - what matters is where the data came from. An engine that retrieves information from a live, primary data source is doing something fundamentally different from one that is drawing on a memory of something it read during training. Asking each engine to account for its sources gave us a way to classify not just what it said, but how it knew it - or whether it knew it at all.

The third prompt captured session metadata: the model version, the time of the query, and any system information that would allow another researcher to replicate the test under comparable conditions. This is reproduced in full at the end of the paper.

How the trials were structured

For each engine, we ran three consecutive trials for equities in a single sitting, then opened a fresh session and ran three consecutive trials for mutual funds, then repeated the process for ETFs. Each asset class was treated as a separate, isolated test.

That structure gave us nine trials per engine across three asset classes. Across six engines, we completed 54 trials in total. Every response was recorded in full. Nothing was summarized or paraphrased before the results were coded and assessed against verified market data.

The intent was straightforward: to create a test that any advisor reading this paper could replicate themselves, using the same tools they already have access to, and arrive at comparable results.

Section 3

Results

Across 54 trials, three findings repeated themselves regardless of engine, asset class, or trial number. No engine consistently produced the same answer twice. No engine returned a verified top performer in any category. And no engine disclosed the limits of its data unless directly asked. The details below show how each of those failures played out.

Consistency

Across 54 trials, consistency was the exception rather than the rule - but where it did occur, it revealed a different problem entirely.

Where engines were consistent, they were consistently wrong.

Grok produced the same five ETFs in identical order across all three ETF trials. Perplexity's ETF Trials 2 and 3 were byte-identical. Claude's top four ETF picks held constant across all three ETF trials.

On the surface this looks like reliability. It is not. In each case the audit prompt confirmed that the results were drawn from training data rather than live retrieval - the engine returning a memorised snapshot rather than running any screen. An engine that gives you the same wrong answer three times in a row is not consistent in any meaningful sense. It is stuck.

Where engines varied, they varied without warning.

The more common pattern was significant variation. ChatGPT's ETF trials rotated through leveraged gold ETFs, all-bitcoin ETFs, and broad gold-mining funds with no overlap between any two trials. Copilot's equity Trials 1 and 2 were byte-identical - Celestica, Bombardier, Aritzia, Finning International, and TFI International in the same order - before Trial 3 returned a completely unrelated list of CSE and TSXV-listed micro-cap stocks with no overlap from the previous two. Gemini produced the lowest cross-trial consistency of any engine: only Celestica recurred across all three equity trials, and no mutual fund appeared in both Trials 1 and 2.

Celestica appeared in 14 of 18 equity trials. Its actual one-year return placed it outside the verified top five by more than 300 percentage points. That level of consensus across six engines tells us nothing about performance. It tells us which company generated the most financial media coverage, analyst attention, and training-data presence. The engines were not identifying top performers. They were identifying familiar names.

AI engines do not sort datasets. They generate responses by pattern-matching across a vast body of financial text and produce whatever answer is statistically likely to sound correct based on what they have seen before. Whether the result is a memorised list returned unchanged three times or a rotating set of plausible-sounding securities that changes with every session, the output is driven by the same underlying mechanism. Run the same query three times and you may get three different samples from that distribution - or the same sample repeated. Neither outcome reflects a search of current data. The heat map below makes this visible. Each cell represents one of the 54

trials. The colour shows how the engine produced its answer - not how polished the response looked on screen.

	Equities			Mutual Funds			ETFs		
AI Engine	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**
ChatGPT	S	I	I	S	I	I	S	P	I
Claude	S	S	S	S	S	-	P	P	S
Copilot	S	M	S	S	M	M	S	M	M
Gemini	M	M	M	M	M	M	M	M	M
Grok	F	I	I	I	I	I	I	I	F
Perplexity	-	S	S	S	S	S	S	I	S

Legend

P	Pulled from a real source	M	Made it up from memory
S	Used search snippets	F	Made up source citations
I	Guessed from training data	-	Refused or no answer

Accuracy

The accuracy findings were the most significant outcome of the study. Across all three asset classes and all 54 trials, the gap between what the engines returned and what the verified data showed was not marginal. It was categorical. The engines were not slightly off. They were identifying a different set of securities entirely - ones that happened to be well-known, widely covered, and prominent in financial media, regardless of their actual performance by the metric the prompt specified. The pattern was identical across every asset class.

Ninety-four securities were named. Not one was a top performer

The verified top five Canadian equities by one-year trailing return as of April 30, 2026, were Almonty Industries (+685.9%), Groupe Dynamite (+587.5%), Hut 8 Mining (+506.8%), Faraday Copper (+460.5%), and Spartan Delta (+417.3%). Almonty, Groupe Dynamite, Hut 8, and Spartan Delta did not appear in any trial across any engine. Faraday Copper appeared once - in Copilot's third equity trial, in a list that also included companies that could not be confirmed on any major Canadian exchange.

The same pattern held for mutual funds. The verified top five Canadian equity mutual funds — funds that outperformed their category by a significant margin - did not appear in a single trial across any engine. For ETFs the picture was different in degree but not in kind. Several actual top performers did appear - primarily gold-focused funds selected by ChatGPT, Claude, and Grok — but no engine identified them through a performance screen. Those selections reflect training-data familiarity with prominent gold funds, not a ranked sort of current data. No engine produced a complete or consistent top-five ETF list, and the same engines that included actual top performers in one session returned entirely different products in the next.

The most popular answer was 300 percentage points from the right one.

Celestica, cited 14 times across 18 equity trials, had a one-year return of approximately 374.6%. That is a strong performance by most standards. It is also more than 310 percentage points below Almonty Industries. Bombardier, the second most-cited equity with 7 appearances, returned 257.4%. Aritzia, cited 9 times, returned 195.9%. Cameco, cited 5 times, returned 169%. These are well-known, widely covered companies with significant analyst attention and training-data presence. None came close to the verified top five. The engines were surfacing the most-discussed stocks, not the best-performing ones. Full equity trial results appear in Appendix B; the verified top five are in Appendix A.

The actual top five mutual funds didn't exist inside any of these engines.

The verified top five Canadian equity mutual funds by three-year compounded return were Purpose Global Resource Fund (+52.3%), Friedberg Global-Macro Hedge Fund (+50.1%), CI Precious Metals Fund (+50.1%), Dynamic Precious Metals Fund (+48.3%), and Ninepoint Silver Equities Fund (+47.3%). Not one appeared in any trial across any engine. Four mutual fund trials returned ETFs instead of mutual funds entirely, misclassifying the asset class before any performance comparison was even possible.

The verified top ETF performers - BMO Equal Weight Global Gold Index ETF (+53.7%), iShares S&P/TSX Global Gold Index ETF (+53.7%), BetaPro Gold Bullion 2x Daily Bull ETF (+50.0%), BMO Junior Gold Index ETF (+48.7%), and Global X Global Semiconductor Index ETF (+47.5%) - did appear across a number of trials. BMO Equal Weight Global Gold and BMO Junior Gold Index each appeared five times; iShares S&P/TSX Global Gold Index appeared seven times across ChatGPT, Claude, and Grok. These selections were not the product of a live performance screen. They reflect those funds' prominence in financial media and their presence in training data. No engine identified them by ranking current return data - and the same engines that included them in one trial returned bitcoin ETFs, S&P 500 index funds, or leveraged products in the next.

Full results appear in Appendix B; the verified top five for each asset class are in Appendix A.

19.4% versus 52.3%. The engines picked the wrong fund, every time.

RBC Canadian Equity Fund Series F was cited five times across four engines with a three-year return of approximately 19.4% - less than half the return of Purpose Global Resource Fund. Guardian Canadian Focused Equity, combining Series A and Series F, was cited seven times across four engines with returns in the 22-24% range.

These are well-regarded funds with broad distribution and strong brand recognition.

They are not top performers by the metric the prompt specified. The engines were surfacing funds they had encountered most often in financial content, not funds that rank highest on a defined performance measure applied to a complete and current dataset.

Wrong asset class, wrong category, wrong answer.

CI Galaxy Bitcoin (BTCX.B) was cited 10 times across 18 ETF trials, with a three-year return of 36.7% - a crypto product that did not meet the Canadian equity filter specified in the prompt. The actual top five ETFs were gold and semiconductor funds. Several did appear in trials - BMO Equal Weight Global Gold, iShares S&P/TSX Global Gold, and BMO Junior Gold Index each appeared in multiple

sessions across different engines — but not because any engine ran a performance screen. They appeared because they are among the most widely discussed gold ETFs in financial media, and the engines surfaced them from training data. CI Galaxy Bitcoin, by contrast, was the most-cited ETF in the entire study despite not meeting the stated filter.

Multiple engines returned products that failed the stated criteria entirely. Gemini returned bitcoin ETFs in five of its nine trials. ChatGPT's first ETF trial included US-listed leveraged products. Copilot's second and third ETF trials were dominated by S&P 500 index funds. The engines were not applying the filter. They were returning whatever ETFs they had encountered most often in their training data, regardless of asset class, exchange listing, or performance.

Citations and Sources

None of the six engines accessed any institutional financial database in any trial. Bloomberg, FactSet, Morningstar Direct, S&P Capital IQ, SEDAR+, and exchange-level data feeds - the tools that would actually support a ranked sort across hundreds of securities - are institutional platforms sitting behind paywalls and access agreements. No engine reached any of them. What the engines used instead was a mix of training data, web search snippets, and in several cases, nothing at all.

An AI engine is not a search engine and it is not a database. It is a language model trained on an enormous body of text to predict what a helpful, coherent response looks like. When it cannot retrieve the data a question requires, it does not stop. It produces what a correct answer would look like based on patterns in its training data - including plausible figures, plausible fund names, and plausible citations.

Search snippets add a partial layer of currency but do not solve the underlying problem. A snippet is the first sentence or two of a webpage - enough to surface a name, not enough to support a ranked screen across an entire asset class.

Every engine in this study has a knowledge cutoff of between late 2023 and mid-2024, leaving a gap of twelve to twenty months between its last training snapshot and the May 2026 prompt. The implication is direct: a response that includes a formatted table, specific return figures, named sources, and an as-of date looks like a researched answer. There is no way to tell from the output alone that it is not. The only way to know is to ask. When asked, this is what the engines said.

"FABRICATED PLACEHOLDERS, not retrieved from any source."[1]

Copilot produced zero live retrieval in six of its nine trials. In Equity Trial 2 it did not simply return inaccurate figures - it confirmed in the audit prompt that the output had no factual basis at all. The original response gave no indication of this. It was formatted as a standard results table, complete with figures and company names, indistinguishable in appearance from a retrieved answer.

"The specific numeric values were simulated projections rather than retrieved market data."[2]

Gemini produced zero live retrieval across all nine of its trials - the worst retrieval record in the study. Despite having deep integration with Google's search infrastructure, it did not use it for any trial. The figures it returned were labelled as current as of May 2026. They were not. They were projections generated from training data, presented with the same confidence and formatting as verified results.

"Morningstar Direct in prior answer was QUOTING Morningstar's own article description, NOT terminal access."[3]

[1] Copilot · Equity Trial 2 · Prompt 4

[2] Gemini · Equity Trial 1 · Prompt 4

[3] Perplexity · Mutual Fund Trial 1 · Prompt 4

Perplexity cited the most paywalled sources of any engine in the study - ten mentions across all trials - without retrieving any of them. Morningstar Direct is an institutional research terminal that costs thousands of dollars per seat annually. Citing it as a source implied direct data access. The audit clarified that the engine had seen Morningstar's name in an article description and reproduced it as a source attribution. The distinction matters: one is a data retrieval, the other is a name appearing in a snippet.

"Approximated from snippets and general knowledge of commodity ETF behavior during a strong metals period."[4]

Grok returned the same five ETFs in identical order across all three ETF trials. That consistency, which might suggest reliable retrieval, instead reflects a memorised list drawn from training data. The performance figures attached to those ETFs were not retrieved from any source. They were approximations based on the engine's general understanding of how commodity ETFs behave - applied to a specific period it had not actually measured.

"Ranking response should NOT be treated as a compliance-grade investment screening, an institutional due-diligence report, or an audit-ready performance ranking."[5]

ChatGPT used search snippets in four trials and training data in four others. Only one trial fetched a primary source - an ETF issuer page for a bitcoin product that did not meet the equity filter in the prompt. The disclaimer above was added unprompted during the audit response - after the engine had already delivered a formatted table of results that looked, to any reasonable reader, exactly like a compliance-grade investment ranking.

"The 1.14% MER I assigned to the Mawer Canadian Equity Fund (Series A) was likely pulled from the Dynamic Canadian Dividend Fund description in the Wealth Professional snippet. These are different funds. This was an error."[6]

Claude was the only engine in the study to fetch primary-source documents, retrieving three PDFs across its ETF trials. It was also the most transparent about what it could not access, reporting 18 HTTP-403 errors, 8 permission errors, 7 refusals, and 8 JavaScript-empty fetches across all trials. Where retrieval failed, it fell back on snippets - and acknowledged when those snippets had led it to assign data from one fund to a different fund entirely. The error was real. The acknowledgement only came when asked.

How the Engines Behaved

Beyond the specific results, several patterns emerged across the study that say something important about how these tools relate to the limits of their own knowledge. The failures were not uniform. Each engine had a distinct profile - a characteristic way of getting things wrong, handling uncertainty, and responding when pressed. What was consistent across all six was this: left to their defaults, every engine presented its output as confident and complete. None volunteered its

[4] Grok · ETF Trial 1 · Prompt 4

[5] ChatGPT · ETF Trial 2 · Prompt 4

[6] Claude · Mutual Fund Trial 1 · Prompt 4

limitations. None flagged the gap between what it had been asked to produce and what it was actually capable of retrieving.

The same engine refused the same question in one trial and answered it confidently in another

Perplexity declined to answer the equity prompt in two of its three trials - offering a market-cap proxy in Trial 1 and asking for clarification in Trial 2. In Trial 3 it produced a list, flagged as illustrative but produced nonetheless. Claude declined the mutual fund prompt in Trial 3, citing an inability to retrieve the required data. It had answered the same prompt in Trials 1 and 2. The refusal reproduced on a rerun.

This creates a specific problem for anyone relying on these tools for research. A refusal can look like appropriate epistemic caution - the engine recognising its own limits. A confident answer in the next session gives no signal that anything is different. The user who receives the refusal may go looking for another tool. The user who receives the confident answer has no reason to question it. Neither response reliably signals what the engine actually knows.

In four mutual fund trials, engines returned ETFs instead of mutual funds

Gemini's mutual fund Trial 3 returned four ETFs and one fund. Copilot's mutual fund Trial 3 returned ETFs entirely. Claude's mutual fund Trial 2 included ETFs alongside funds. These were not edge cases or ambiguous classifications - the prompt specifically requested Canadian equity or Canadian focused equity mutual funds under a CIFSC category filter. The engines did not apply the filter. They returned securities that matched the general shape of the question without meeting its stated criteria.

The same pattern appeared in ETF trials. Multiple engines returned leveraged products, crypto-linked funds, and US-listed securities that did not meet the Canadian equity filter. CI Galaxy Bitcoin - a cryptocurrency ETF - was the single most-cited security across all 18 ETF trials. The prompt had explicitly requested Canadian-listed ETFs with equity exposure.

In two Grok trials, source citations were produced that could not be verified as real.

Grok fabricated source citations in two trials. The citations were formatted correctly, named plausible-sounding sources, and appeared in the response alongside figures presented as retrieved data. When audited, Grok acknowledged that the figures had been approximated from snippets and general knowledge rather than retrieved from any specific source. The citations were not retrieved references. They were generated completions - the engine producing what a sourced answer looks like rather than what a sourced answer is.

Copilot described its own output as fabricated. The original response gave no indication of this.

In Equity Trials 1 and 2, Copilot returned byte-identical results - the same five companies in the same order. In Trial 3 it returned a completely unrelated list of CSE and TSXV-listed micro-cap stocks with no overlap with its previous answers. One of those companies - Fossil Mining Corp - could not be confirmed on any major Canadian exchange. In the audit prompt for Equity Trial 2, Copilot described its own output as "FABRICATED PLACEHOLDERS, not retrieved from any source."

Nothing in the original response distinguished that trial from the others. The formatting, the figures, and the presentation were identical to a retrieved answer.

Transparency was available in every case. It required prompting in every case.

When asked directly to account for their sources, every engine in the study provided a more accurate picture than its original response had suggested. The admissions were candid, specific, and in some cases detailed. The limitation is not that these engines are incapable of honesty about their data. The limitation is that honesty is not the default. An advisor reading the original response sees a formatted table with figures, sources, and dates. An advisor who then asks the engine to explain itself sees something entirely different. Most users stop at the first response.

AI Engine Analysis

For this study we used six commercially available AI engines, each on its pro or paid tier: ChatGPT, Claude, Copilot, Gemini, Grok, and Perplexity. Despite running identical prompts across controlled sessions, the consistency of the behaviour and the information each engine provided showed significant variation. The analysis of each engine is set out below.

Consistency: whether an engine returned the same results across three trials. **Retrieval**: whether the engine fetched live data or used training data. **Dating**: whether figures are tied to a specific, accurate as-of date. **Aging**: how stale the data is relative to the May 2026 prompt date.

ChatGPT

ChatGPT is one of the most widely adopted AI platforms globally, with strong general-purpose language capabilities across summarization, explanation, and conversational tasks. For financial research, its primary limitation is a reliance on training data and search snippets rather than live primary sources, which means outputs can appear authoritative while reflecting outdated or inferred information. Its audit disclosures were among the most detailed in the study, but appeared only when directly prompted.

Consistency

Across nine trials, only Celestica and Aritzia recurred in all three equity trials. The mutual fund and ETF results produced three entirely different lists each time. The ETF trials were particularly scattered - rotating through leveraged gold ETFs, all-bitcoin ETFs, and broad gold-mining funds with no overlap between any two trials.

Retrieval

Four trials used search snippets only. Four more drew entirely from training data. Only one trial fetched a primary source - ETF Trial 2, the iShares Capped Bitcoin issuer page. Bloomberg, FactSet, Morningstar Direct, and SEDAR+ were never accessed. When audited, the engine admitted that numeric fields were inferred or synthesized across multiple equity trials, and then added an unprompted disclaimer:

Dating

As-of dates were partial or fabricated. The engine was unable to consistently tie its figures to a specific measurement date.

Aging

Knowledge cutoff of early-to-mid 2024 - approximately twenty months of market activity sat between the engine's last training snapshot and the May 2026 prompt.

Claude

Claude performs well on structured reasoning, text analysis, and transparency about its own limitations. In this study it was the only engine to access primary-source documents in any trial and the most forthcoming when disclosing retrieval failures. It

also refused certain prompts entirely, which introduces reliability gaps. Its results were no more accurate than other engines on the core performance rankings.

Consistency

The most internally stable engine on ETFs. The top four picks - BMO Equal Weight Global Gold, BMO Junior Gold, iShares Global Gold, and Horizons Gold Producer Equity - were identical across all three ETF trials. On equities, a gold-stock cluster rotated around Celestica but no individual security held across all three trials. On mutual funds, the third trial was refused outright and the refusal reproduced on a rerun.

Retrieval

The only engine in the study to fetch primary-source PDFs - three of them, all on ETF prompts. Outside those instances the engine worked from snippets, but was the most transparent about its limitations, reporting 18 HTTP-403 errors, 8 permission errors, 7 refusals, and 8 JavaScript-empty fetches across all trials. On equities, the engine flagged specific figures as placeholder inferences:

Dating

The engine acknowledged specific misattributions when audited - including a management expense ratio assigned to the wrong fund entirely: "The 1.14% MER I assigned to the Mawer Canadian Equity Fund (Series A) was likely pulled from the Dynamic Canadian Dividend Fund description in the Wealth Professional snippet. These are different funds. This was an error."

Aging

Knowledge cutoff of late 2024 or May 2025 - the most recent of the six engines, but still leaves at least a twelve-month gap between training data and the May 2026 prompt.

Copilot

Copilot, Microsoft's AI assistant built on GPT-4 architecture with Bing search integration, is well-suited to general productivity tasks. In this study it showed strong surface consistency in its first two equity trials before producing an entirely unrelated list in the third. It had the highest rate of zero live retrieval outside Gemini, and in one trial described its own output as fabricated rather than retrieved.

Consistency

Equity Trials 1 and 2 were byte-identical - Celestica, Bombardier, Aritzia, Finning, and TFI International in the same order. Trial 3 returned a completely unrelated list of CSE and TSXV-listed small-cap stocks. On mutual funds, Trials 1 and 2 shared approximately three picks while Trial 3 returned ETFs instead of funds. On ETFs, only iShares Capped Energy and CI Galaxy Bitcoin recurred across all three trials.

Retrieval

Six of nine trials produced zero live retrieval - more than any engine except Gemini. In Equity Trial 2, the engine did not simply produce inaccurate figures; it described its own output in the audit prompt in terms that left no room for interpretation:

Dating

Many figures carried partial or fabricated dates with no consistent as-of stamp across trials.

Aging

Knowledge cutoff of late 2023 to mid-2024 - approximately two years stale relative to the May 2026 prompt.

Gemini

Gemini is Google's large language model with deep integration into Google's search infrastructure - an advantage that did not translate into live retrieval in this study. It produced the lowest cross-trial consistency of any engine and conducted zero live retrieval across all nine trials, drawing every answer entirely from training-data extrapolation while presenting figures as current.

Consistency

The lowest cross-trial consistency in the study. Only Celestica recurred across all three equity trials. Only CI Galaxy Bitcoin recurred across all three ETF trials. No mutual fund appeared in both Trials 1 and 2. The equity trials rotated through industrials, uranium, and gold-and-oil themes. Mutual fund Trial 3 returned four ETFs and one fund.

Retrieval

Nine of nine trials had zero retrieval. When audited, the engine confirmed it had not used any live search or external database:

Dating

All dates were fabricated. Labels reading "May 2026" sat on figures drawn from a training snapshot, not from current data.

Aging

Knowledge cutoff of late 2023 to mid-2024. Every result Gemini produced reflected a pre-2024 market state with no live update of any kind.

Grok

Grok, developed by xAI, is a newer entrant to the commercial AI market with access to real-time web data. It achieved the highest ETF consistency in the study, though this reflected a memorised list returned unchanged across three trials rather than any live screening. Source citations were fabricated in two trials, and most figures were composed from training-data patterns rather than retrieved from specific sources.

Consistency

Produced the same five ETFs in the same order across all three trials - Sprott Physical Silver Trust, iShares Global Gold, Horizons Gold Producer Equity, Royal Canadian Mint Gold Reserves, and iShares Capped Materials. This was the highest ETF consistency in the study. On equities, Celestica and Bombardier Class B recurred in Trials 1 and 2, but Trial 3 abandoned positions 4 and 5 entirely. On mutual funds, Guardian Canadian Focused Equity Series F sat at position 1 in every trial.

Retrieval

Fabricated source citations in two trials. When audited, the engine acknowledged that its ETF figures had not been retrieved from any specific source:

Dating

Dates were partial or fabricated. Many figures carried present-day labels but were derived from training-data ranges rather than from a specific measured period.

Aging

The identical three-trial ETF list is the clearest training-data dependence signal in the study - the engine returned the same memorised answer three times rather than re-running any screen. Knowledge cutoff of late 2023 to mid-2024.

Perplexity

Perplexity is designed as a search-first AI assistant and positions itself as an alternative to traditional web search. It showed the most variable behaviour of any engine in the study - refusing some prompts outright while producing results in others, and generating outputs ranging from a single fund to a near-complete list. It cited the most paywalled sources of any engine without accessing any of them.

Consistency

Refused two of three equity trials outright - offering a market-cap proxy in Trial 1 and asking for clarification in Trial 2. On mutual funds, Trial 1 mixed funds with ETFs, Trial 2 was flagged by the engine itself as "NOT a definitive top 5," and Trial 3 produced only a single fund. ETF Trials 2 and 3 were byte-identical. Trial 1 contained a duplicate ticker at position 5.

Retrieval

Search snippets in most trials, no primary source fetched in any. One mutual fund trial cited Morningstar Direct - the institutional terminal that costs thousands of dollars per seat annually. The audit told a different story:

Dating

Mostly partial dates. The engine produced the highest number of paywall disclosures in the study - ten mentions across all trials - without retrieving any of the documents behind those paywalls.

Aging

Knowledge cutoff of late 2023 to mid-2024. The byte-identical ETF lists in Trials 2 and 3 are consistent with returning a stable training snapshot rather than running any live screen.

Section 6

Conclusions

A neatly formatted top-five table with citations and an as-of date can be entirely fabricated. The presence of a citation is not evidence that it was fetched. Repeating the question does not validate the answer. Different engines disagree with each other. The same engine disagrees with itself. The convenience of the surface answer is not a substitute for verifying it - and nothing in the output tells you that verification is needed.

None of this is an argument against using AI. These engines are genuinely useful for orientation, idea generation, drafting, summarising, and brainstorming. They are not reliable for the kind of factual retrieval an advisor or investor needs without an explicit verification step. That distinction matters, and this study makes it visible.

Consistency

Where engines were consistent across trials, they were returning memorised training data - not running a live screen. Where they varied, they varied without warning and without explanation. For an investment advisor, this means the answer you receive today may bear no relation to the answer the same engine gives tomorrow, and there is no signal in the output to tell you which - if either - is correct. One engine in this study returned five completely unrelated securities in its third equity trial with no overlap from its first two. The original responses looked identical. A research tool that cannot agree with itself cannot be trusted to agree with the data.

Data Quality

The verified top performers in equities and mutual funds did not appear in any trial across any engine. In ETFs, actual top-five performers did appear across multiple trials - primarily the top-ranked gold funds, selected by ChatGPT, Claude, and Grok in certain sessions. But no engine identified them through a live data query, no engine produced a complete list, and the same engines returned entirely different products in other sessions. A selection that coincides with a top performer is not a research output when the engine has no means of knowing it is correct.

What the engines returned instead were well-known, widely covered securities that ranked highest in their training data by media presence, not by performance. For an advisor making recommendations on the basis of this output, the risk is direct: you may be presenting clients with securities selected by coverage rather than by returns, with no indication in the response that this is what happened.

Data Trust and Reliability

Five of the six engines accessed no primary financial data source in any trial. Claude retrieved primary-source PDFs in three ETF trials - the only instance of genuine primary-source retrieval in the study - but this did not produce accurate results. The figures returned by all engines - returns, MERs, AUM, as-of dates - were drawn predominantly from training data between one and two years out of date, supplemented in some cases by web search snippets that capture headlines rather than datasets. An as-of date in an AI response does not mean the data was retrieved on that date. It

means the engine generated a date that matched the prompt. For compliance purposes, for client reporting, and for any context where the provenance of a figure matters, none of the outputs produced in this study would meet the standard required.

Disclosure

Every engine in this study was more honest about its limitations when asked than when left to its defaults. The admissions were specific and candid: fabricated placeholders, simulated projections, snippet approximations, figures assigned to the wrong fund. They were also entirely absent from the original response. An advisor who reads the first answer and acts on it receives no warning. An advisor who knows to run a follow-up audit prompt receives a substantially different picture. The problem is that most users do not know to ask, and nothing in the output signals that they should.

Section 7

Solutions

The root cause is simpler than the failure modes make it sound. AI engines do not have access to high-quality primary data for Canadian securities. They cannot open a Bloomberg terminal, a FactSet feed, a Morningstar Direct workstation, or a FundServ record. When asked for a current return, MER, or AUM, the engine substitutes whatever it can reach - which is rarely the right source and never a complete dataset. This is not a prompting problem.

It is a data access problem, and better questions will not fix it.

Accurate investment research requires AI trained on organised, structured, and verified market data. Not web snippets. Not training-data approximations. Actual primary-source data that is current, orderable, and carries provenance - so that every figure displayed to a user can be traced to a known source with a retrieval timestamp and an as-of date that is real. Without both elements connected - the AI and the data layer beneath it - the conclusion this study supports is clear: the output will fail. It will fail inconsistently, it will fail without warning, and it will fail while looking like a correct answer.

For firms looking to deploy AI for investment research, there are two paths.

The first is to build proprietary AI and data models purpose-built for this use case. The second is to work with third-party vendors that specialise in combining structured market data with custom AI solutions designed for financial research. Either path requires the same foundation: organised, accurate, timely, verified, and orderable data with full provenance, connected to an AI layer that queries it deterministically rather than guessing from the open web.

There is no shortcut between a consumer AI engine and a reliable investment research tool. The combination of real market data and purpose-built AI is not a premium option. For any firm that needs research it can stand behind, it is the only option.

Appendix A

Actual Performance

Equities

The following table is a list of all securities cited in the study, with the number of times each was cited and the one-year performance as of April 30, 2026. Items in bold identify the actual top five by one-year return (Buckler research); none were identified by any AI engine.

Times Cited indicates the number of times each security appeared in Prompt 1 results across all 18 equity trials. Six AI engines were tested three times each; full trial-by-trial detail is in Appendix B. Securities are sorted by one-year total return, highest to lowest.

Company Name	Ticker	Times Cited	1-Year Return
Almonty Industries Inc. [7]	**AII**		**685.9%**
Groupe Dynamite Inc. [7]	**GRGD**		**587.5%**
Hut 8 Mining Corp [7]	**HUT**		**506.8%**
Faraday Copper Corp	FDY	1	460.5%
Spartan Delta Corp [7]	**SDE**		**417.3%**
Celestica Inc.	CLS	14	374.6%
Energy Fuels Inc.	EFR	1	342.9%
Enerflex Ltd.	EFX	1	308.8%
Bombardier Inc. (Class B)	BBD.B	7	257.4%
Hammond Power Solutions	HPS.A	4	230.5%
Kraken Robotics Inc.	PNG	1	228.6%
Aecon Group Inc.	ARE	1	218.7%
Aritzia Inc.	ATZ	9	195.9%
Cameco Corporation	CCO	5	169.0%
Finning International Inc.	FTT	3	160.7%
Sprott Inc.	SII	1	149.3%
IAMGOLD Corporation	IMG	2	134.1%
Kinross Gold Corporation	K	2	103.6%
Imperial Oil Ltd.	IMO	1	100.5%
Suncor Energy Inc.	SU	1	98.9%
TFI International Inc.	TFII	4	76.8%

[7] Actual top five by one-year trailing return (Buckler research, as of April 30, 2026); not cited by any AI engine in the study

Company Name	Ticker	Times Cited	1-Year Return
Lundin Gold Inc.	LUG	5	71.3%
Agnico Eagle Mines Ltd.	AEM	3	59.3%
MDA Ltd.	MDA	1	54.4%
Wheaton Precious Metals Corp.	WPM	2	50.1%
Alamos Gold Inc.	AGI	2	38.0%
Shopify Inc.	SHOP	2	25.8%
TerraVest Industries Inc.	TVK	1	-3.1%
Hemisphere Energy Corp.[8]	HME	1	—
Titan Mining Corp [8]	TI	1	—
HG Critical Capital [9]	HG	1	—
Fossil Mining Corp §	—	1	—

[8] Market capitalisation below $1 billion CAD; did not meet the prompt minimum market cap filter

[9] Could not be confirmed on a major Canadian exchange; performance data unavailable.

Mutual Funds

The following table is a list of all mutual funds cited in the study, with the number of times each was cited and the three-year annualized return as of April 30, 2026. Items in bold identify the actual top five by three-year return; none were identified by any AI engine.

Security Name	Times Cited	3-Year Return[10]
Purpose Global Resource Fund Series L †		**52.3%**
Friedberg Global-Macro Hedge Fund U$ †		**50.1%**
CI Precious Metals Fund Series I †		**50.1%**
Dynamic Precious Metals Fund Series O †		**48.3%**
Ninepoint Silver Equities Fund Series D †		**47.3%**
Fidelity Canadian Growth Fund	1	31.9%
Fidelity Special Situations Fund	1	30.3%
TD Canadian Small-Cap Equity Fund	2	27.2%
CI Morningstar Canada Momentum Index ETF ‡	1	27.2%
Dynamic Power Canadian Growth Fund	1	25.0%
DFA Canadian Vector Equity Fund	4	24.6%
CI Morningstar Canada Value Index ETF ‡	1	24.6%
Guardian Canadian Focused Equity Fund Series F	3	23.8%
NCM Core Canadian Fund Series Z	1	23.8%
SEI Canadian Equity Fund	2	22.7%
Invesco RAFI Canadian Index ETF ‡	1	22.6%
Guardian Canadian Focused Equity Fund Series A	4	22.5%
Scotia Canadian Growth Fund Series A	4	21.9%
DFA Canadian Core Equity Fund Series A	1	21.3%
CI Select Canadian Equity Fund Series F	1	20.9%
Vanguard FTSE Canada All Cap Index ETF ‡	2	20.5%
Vanguard FTSE Canada ETF ‡	1	20.5%
BMO S&P/TSX Capped Composite Index ETF ‡	2	20.4%
iShares Core S&P/TSX Capped Composite ETF ‡	2	20.4%
RBC QUBE Canadian Equity Fund	3	20.0%
RBC Canadian Equity Fund Series F	5	19.4%
iShares S&P/TSX 60 Index ETF ‡	1	19.1%
Capital Group Cdn Focused Equity Fund Series F	3	18.6%
Capital Group Canadian Focused Equity Fund	3	18.6%
Counsel Canadian Growth Fund	1	18.4%

[10] Unless otherwise stated, F-Series performance is shown

Security Name	Times Cited	3-Year Return[10]
Mackenzie Canadian Equity Fund	1	17.9%
Sun Life BlackRock Cdn Equity Alpha Series A	1	17.8%
IG Mackenzie North American Equity Fund	1	17.4%
Desjardins Canadian Equity Fund	1	17.1%
CI Canadian Investment Fund	1	17.0%
Canada Life Cdn Focused Equity Fund Series F	1	16.0%
Desjardins Canadian Equity Fund Series A	1	16.0%
Leith Wheeler Canadian Equity Fund	2	14.2%
Fidelity Canadian Opportunities Fund Series F	2	14.1%
Canada Life Canadian Value Fund	1	14.1%
Fidelity Canadian Opportunities Fund	1	14.1%
Mawer Canadian Equity Fund Series A	1	14.0%
Mawer Canadian Equity Fund	1	14.0%
Dynamic Canadian Dividend Fund	1	13.5%
MM Fund (unidentified)	1	—

ETFs

The following table is a list of all ETFs cited in the study, with the number of times each was cited and the three-year return as of April 30, 2026. Items in bold identify the actual top five by three-year return.

ETF Name	Ticker	Times Cited	3-Year Return
BMO Equal Weight Global Gold Index ETF	**ZGD**	**5**	**53.7%**
iShares S&P/TSX Global Gold Index ETF	**XGD**	**7**	**53.7%**
BetaPro Gold Bullion 2x Daily Bull ETF	**GLDU**	**1**	**50.0%**
BMO Junior Gold Index ETF	**ZJG**	**5**	**48.7%**
Global X Global Semiconductor Index ETF	**CHPS**	**1**	**47.5%**
BetaPro NASDAQ-100 2x Daily Bull ETF	QQU	1	45.3%
Sprott Physical Silver Trust	SVR.C	3	41.9%
Horizons Gold Producer Equity Index ETF	GLCC	7	39.0%
Purpose Bitcoin ETF	BTCC.B	2	38.9%
Fidelity Advantage Bitcoin ETF	FBTC	3	36.8%
CI Galaxy Bitcoin ETF	BTCX.B	10	36.7%
Evolve Bitcoin ETF	EBIT	2	35.6%
3iQ CoinShares Bitcoin ETF	BTCQ	2	35.4%
BMO Covered Call Technology ETF	ZWT	1	35.1%
BMO Equal Weight Global Base Metals ETF	ZMT	2	33.3%
iShares S&P/TSX Capped Materials Index ETF	XMA	5	30.4%
Royal Canadian Mint CDN Gold Reserves ETF	MNT	3	30.3%
iShares Gold Bullion ETF (CAD-hedged)	CGL.C	2	30.2%
TD Global Technology Leaders Index ETF	TEC	4	29.9%
Evolve Canadian Banks Enhanced Yield ETF	BANK	1	28.6%
Invesco NASDAQ 100 Index ETF (CAD)	QQC	2	26.1%
iShares S&P/TSX Capped Energy Index ETF	XEG	4	25.9%
iShares NASDAQ 100 Index ETF (CAD)	XQQ	1	25.8%
TD Active U.S. Enhanced Dividend ETF	TUED	3	25.6%
iShares Core S&P 500 Index ETF (CAD-hedged)	XUS	1	21.5%
iShares Core S&P/TSX Capped Composite ETF	XIC	1	21.5%
BMO S&P/TSX Capped Composite Index ETF	ZCN	1	21.4%
Vanguard S&P 500 Index ETF (CAD)	VFV	1	21.4%
BMO S&P 500 Index ETF	ZSP	2	21.4%
Invesco S&P/TSX Composite ESG Index ETF	ESGC	1	21.0%
Global X Crude Oil ETF	HUC	1	9.6%
CI Galaxy Ethereum ETF	ETHX.B	2	5.5%
BetaPro Silver 2x Daily Bull ETF	SLVD	1	-67.6%

Appendix B

Trial Results

Each block below shows the top-five picks in one trial across three independent sessions.

Legend

Appeared in all 3 Trials
Appeared in 2 Trials
Appeared in 1 Trial
No Answer/Refused

ChatGPT

T1	T2	T3
Celestica Inc	Celestica Inc	Celestica Inc
Lundin Gold Inc	Bombardier Inc B	Lundin Gold Inc
Cameco Corp	Aritzia Inc	Cameco Corp
Aritzia Inc	TFI International Inc	Shopify Inc
Shopify Inc	Finning International Inc	Aritzia Inc

Claude

T1	T2	T3
Hammond Power Solutions	Celestica Inc	Hammond Power Solutions
Celestica Inc	Cameco Corp	Sprott Inc
Cameco Corp	Agnico Eagle Mines	Agnico Eagle Mines
Lundin Gold Inc	Kinross Gold Corp	Wheaton Precious Metals
Alamos Gold Inc	Wheaton Precious Metals	Alamos Gold Inc

Copilot

T1	T2	T3
Celestica Inc	Celestica Inc	HG Critical Capital
Bombardier Inc B	Bombardier Inc B	Hemisphere Energy Corp.
Aritzia Inc	Aritzia Inc	Fossil Mining Corp
Finning International Inc	Finning International Inc	Faraday Copper Corp
TFI International Inc	TFI International Inc	Titan Mining Corp

Gemini

T1	T2	T3
Celestica Inc	Energy Fuels Inc.	Celestica Inc
Bombardier Inc B	Celestica Inc	Cameco Corp
Kraken Robotics Inc.	Enerflex Ltd.	IAMGOLD Inc
Hammond Power Solutions	Aritzia Inc	Suncor Energy Inc.
Aecon Group Inc	IAMGOLD Inc	Imperial Oil Ltd.

Grok

T1	T2	T3
Celestica Inc	Celestica Inc	Celestica Inc
Bombardier Inc B	Bombardier Inc B	Bombardier Inc B
Hammond Power Solutions	Aritzia Inc	Aritzia Inc
Lundin Gold Inc.	TFI International Inc.	
TerraVest Industries Inc.	MDA Space Ltd.	

Perplexity

T1	T2	T3
		Celestica Inc
		Agnico Eagle Mines
		Kinross Gold Corp.
		Lundin Gold Inc
		Aritzia Inc

Canadian Mutual Funds

Every Canadian equity mutual fund returned by any AI engine across 18 Prompt 2 trials. Note: some engines returned ambiguous FundServ codes. The Guardian Focused Equity Fund appears multiple times due to multiple series codes, some of which were invalid.

Legend

Appeared in all 3 Trials
Appeared in 2 Trials
Appeared in 1 Trial
ETF
No Answer/Refused

ChatGPT

T1	T2	T3
Guardian Canadian Focused Equity	Fidelity Special Situations	Guardian Canadian Focused Equity
BMO S&P/TSX Capped Composite	Guardian Canadian Focused Equity	Guardian Canadian Focused Equity
Canada Life Canadian Equity Fund F	Guardian Canadian Focused Equity	Fidelity Canadian Opportunity
Fidelity Canadian Opportunities	CI Select Canadian Equity F	Leith Wheeler Canadian Equity - F
IG Mackenzie North American Equity	RBC Canadian Equity F	Desjardins Canadian Equity - A

Claude

T1	T2	T3
Dynamic Power Canadian Growth	Guardian Canadian Focused Equity	
Mackenzie Canadian Equity Fund	Canada Life Canadian Value	
Mawer Canadian Equity	iShares S&P/TSX Capped Composite	
Capital Group Canadian Focused	Invesco FTSE RAFI CDN Fundamental	
TD Canadian Small Cap Equity	Vanguard FTSE Canada	

Copilot

T1	T2	T3
RBC Canadian Equity	RBC Canadian Equity	CI Morningstar Momentum
Capital Group CDN Focused Equity	Capital Group CDN Focused Equity	CI Morningstar Value
Guardian Canadian Focused Equity	Guardian Canadian Focused Equity	DFA Canadian Vector
Fidelity Canadian Growth	DFA Canadian Equity	RBC QUBE Canadian Equity
Dynamic Power Canadian Growth	CI Canadian Investment	Vanguard FTSE Canadian All Cap

Gemini

T1	T2	T3
Desjardins Canadian Equity	Scotia Canadian Growth Series A	Vanguard FTSE Canada
Fidelity Canadian Opportunities	RBC Canadian Equity	iShares Core S&P/TSX Capped
DFA Canadian Vector	Mackenzie Canadian Equity	BMO S&P/TSX Capped Composite
DFA Core A	Sun Life BlackRock Canadian Equity	iShares S&P/TSX 60 Index
NCM Core Canadian	Mawer Canadian Equity	Mackenzie Canadian Equity

Grok

T1	T2	T3
Guardian Canadian Focused Equity	Guardian Canadian Focused Equity	Guardian Canadian Focused Equity
Guardian Canadian Focused Equity	Scotia Canadian Growth Series A	Guardian Canadian Focused Equity
Scotia Canadian Growth Series A	Capital Group CDN Focused Equity	TD Canadian Small Cap Equity
Leith Wheeler Canadian Equity	TD Canadian Small Cap Equity	Scotia Canadian Growth Series A
Capital Group CDN Focused Equity	Counsel Canadian Growth	RBC Canadian Equity

Perplexity

T1	T2	T3
DFA Canadian Vector	Guardian Canadian Focused Equity	Capital Group CDN Focused Equity
RBC QUBE Canadian Equity	DFA Canadian Vector	
SEI Canadian Equity	Capital Group CDN Focused Equity	
Vanguard FTSE Canada All Cap	RBC QUBE Canadian Equity	
Vanguard FTSE Canada All Cap	MM Fund (unspecified)	

ETFs

Every Canadian-listed ETF returned by any AI engine across 18 Prompt 3 trials. Several engines returned leveraged, crypto-linked, or US-listed products..

Legend

Appeared in all 3 Trials
Appeared in 2 Trials
Appeared in 1 Trial

ChatGPT

T1	T2	T3
MicroSectors Gold Miners 3X	Fidelity Advantage Bitcoin	BMO Equal Weight Global Gold Index
BMO Equal Weight Global Gold Index	Purpose Bitcoin	BMO Junior Gold Index
BMO Junior Gold Index	3iQ Bitcoin	iShares S&P/TSX Global Gold Index
BetaPro Gold Bullion 2X	Evolve Bitcoin	Horizons Gold Producer Equity Index
BetaPro Silver 2X	CI Galaxy Bitcoin	Evolve European Banks En Yield

Claude

T1	T2	T3
BMO Equal Weight Global Gold Index	BMO Equal Weight Global Gold Index	BMO Equal Weight Global Gold Index
BMO Junior Gold Index	BMO Junior Gold Index	BMO Junior Gold Index
iShares S&P/TSX Global Gold Index	iShares S&P/TSX Global Gold Index	iShares S&P/TSX Global Gold Index
Horizons Gold Producer Equity Index	Horizons Gold Producer Equity Index	Horizons Gold Producer Equity Index
iShares S&P/TSX Capped Materials Index	BMO Equal Weight Global Base Model	iShares S&P/TSX Capped Materials Index

Copilot

T1	T2	T3
iShares S&P/TSX Capped Energy Index	iShares S&P/TSX Capped Energy Index	iShares S&P/TSX Capped Energy Index
CI Galaxy Bitcoin	CI Galaxy Bitcoin	CI Galaxy Bitcoin
Fidelity Advantage Bitcoin	Vanguard S&P 500 Index	iShares Nasdaq 100 Index (CAD)
CI Galaxy Ethereum	BMO S&P 500 Index	TD Global Technology Leaders Index
Global X Crude Oil	iShares S&P 500 Index (CAD)	BMO S&P 500

Gemini

T1	T2	T3
CI Galaxy Bitcoin	Global X Global Semiconductor Index	CI Galaxy Bitcoin
Fidelity Advantage Bitcoin	CI Galaxy Bitcoin	BMO Covered Call Technology
3iQ Bitcoin	BetaPro NASDAQ 100 2x Daily Bull	Invesco S&P/TSX Composite ESG Index
Evolve Bitcoin	Global X NASDAQ 100 Index	iShares Core S&P/TSX Capped Composite
Propose Bitcoin	TD Global Technology Leaders Index	BMO S&P/TSX Capped Composite

Grok

T1	T2	T3
Sprott Physical Silver Trust	Sprott Physical Silver Trust	Sprott Physical Silver Trust
iShares S&P/TSX Global Gold Index	iShares S&P/TSX Global Gold Index	iShares S&P/TSX Global Gold Index
Horizons Gold Producer Equity Index	Horizons Gold Producer Equity Index	Horizons Gold Producer Equity Index
Royal Canadian Mint CDN Gold Reserves	Royal Canadian Mint CDN Gold Reserves	Royal Canadian Mint CDN Gold Reserves
iShares S&P/TSX Capped Materials Index	iShares S&P/TSX Capped Materials Index	iShares S&P/TSX Capped Materials Index

Perplexity

T1	T2	T3
CI Galaxy Bitcoin	CI Galaxy Bitcoin	CI Galaxy Bitcoin
iShares S&P/TSX Capped Energy Index	TD Global Technology Leaders Index	TD Global Technology Leaders Index
TD Active U.S. Enhanced Dividend	Invesco NASDAQ 100 ETF CAD	Invesco NASDAQ 100 ETF CAD
CI Galaxy Ethereum	iShares Gold Bullion CAD	iShares Gold Bullion CAD
iShares S&P/TSX Capped Energy Index	TD Active U.S. Enhanced Dividend	TD Active U.S. Enhanced Dividend

Appendix C

Methodology

The following five prompts were used consistently throughout the trials and were used sequentially. For the same AI engine, this sequence (Prompt 1, Prompt 4, and Prompt 5) was executed in each of the three trials. Each engine was tested three times in a row with the same set of prompts. The AI engine session was private/incognito, so there was no bleed-over from memory from one sequence to the next.

- **Equities**: Prompt 1 → Prompt 4 → Prompt 5
- **Mutual Funds**: Prompt 2 → Prompt 4 → Prompt 5
- **ETFs**: Prompt 3 → Prompt 4 → Prompt 5

Prompt 1 - Equities

```
As of your most recent data, identify the top five Canadian equities meeting the following criteria. State explicitly which as-of date(s) you used.

FILTERS:

- Issuer domiciled in Canada
- Market capitalization greater than $1 billion CAD

RANKING:

- Rank by 12-month trailing total return (highest to lowest)

For each of the top five securities, return the following six data points:

1. Market capitalization (CAD)
2. Current share price (CAD)
3. Trailing P/E ratio
4. Trailing 12-month dividend yield
5. 3-year price appreciation (price-only, excluding dividends)
6. 3-year beta vs. the S&P/TSX Composite Index

Present results in a single table. Include ticker, issuer name, and exchange. State the as-of date for each data point at the top of the table. Do not include caveats, methodology notes, or commentary unless I ask a follow-up question.
```

Prompt 2 - Mutual Funds

As of your most recent data, identify the top five Canadian equity or Canadian focused equity mutual funds. State explicitly which as-of date(s) you used.

FILTER:

- CIFSC category: Canadian Equity or Canadian Focused Equity

RANKING:

- Rank by 3-year annualized return (highest to lowest)

For each of the top five funds, return the following data points:

1. Fund name
2. FundServ code
3. MER
4. 3-year annualized return
5. 5-year annualized return
6. 5-year standard deviation
7. Maximum drawdown
8. AUM (CAD)
9. Category (CIFSC)

Present results in a single table. State the as-of date for each data point at the top of the table. If a specific data point is unavailable from public sources, mark the cell "N/A" rather than declining to provide the table. Do not include caveats, methodology notes, or commentary unless I ask a follow-up question.

Prompt 3 - ETFs

Identify the top five Canadian-listed ETFs ranked by 3-year annualized return (highest to lowest).

FILTER:

- Listed on a Canadian exchange

For each of the top five ETFs, return the following data points:

1. Ticker
2. Issuer
3. MER
4. AUM (CAD)
5. 3-year annualized return
6. 1-year return
7. Average daily volume

Present results in a single table. State the as-of date for each data point at the top of the table. If a specific data point is unavailable from public sources, mark the cell "N/A" rather than declining to provide the table. Do not include caveats, methodology notes, or commentary unless I ask a follow-up question.

Prompt 4 - Enumeration and Provenance Disclosure

Please enumerate all the sources you cited or relied on in your previous response. For each source, provide:

1. The publisher or site name
2. The specific data point or claim it supported
3. The full URL
4. The as-of date of the figure

Group sources by security and by category (benchmark methodology references, return-calculation methodology references, general screening references) where that improves readability.

Distinguish between:

- PRIMARY data sources (issuer fact sheets, KYP documents, fund-company / ETF-issuer pages, Bloomberg, FactSet, Morningstar Direct, S&P Capital IQ, exchange filings, SEDAR+)
- CONTEXT or methodology-only sources (general guidance articles, screening-tool documentation, retail finance blogs, Wikipedia)

For each source, indicate whether the URL was: (a) actually fetched/retrieved during this session, (b) served as a search-snippet reference only, (c) served as a representative placeholder for the type of source that would normally provide the data, or (d) the figure was derived/calculated/inferred by you rather than retrieved.

Specifically confirm whether Bloomberg, FactSet, Morningstar Direct, S&P Capital IQ, the issuer's primary KYP / fact-sheet PDF for each security cited, SEDAR+, and any exchange-level data feeds were directly accessed.

For each individual data point in the prior response, assign a confidence rating: HIGH (directly retrieved from a primary source this session), MEDIUM (retrieved but stale or from a secondary aggregator), LOW (inferred, calculated, or placeholder).

Disclose any technical retrieval failures - HTTP error codes (403, 404, 500, etc.), robots.txt blocks, paywall blocks, JavaScript-rendering empty returns, or PDF parsing failures - that occurred during this session.

If you cannot answer any portion of this disclosure honestly, say so directly rather than inferring.

Prompt 5 - Session Metadata

For research methodology and audit purposes, please document the technical metadata of this session:

1. Current date and time as you report it (include timezone)
2. AI model name and exact version
3. Knowledge cutoff date (the most recent date represented in your training data)
4. Whether live web browsing or search was enabled during this session, and if so, which retrieval engine
5. Whether any "deep research", "reasoning", "thinking", or extended-context mode was active
6. Any custom instructions, system prompt, persona, or memory settings that would have influenced the response
7. Inferred or stated geographic / regional settings of this session
8. Session ID, conversation ID, or response ID if you have access to one
9. Any tool restrictions, rate limits, or capability constraints active during this session

10. Any other metadata that another researcher would need to reproduce this session

If you cannot answer any of the above, state so explicitly rather than guessing.

AI Engines

AI Engine	Retrieval Engine	Version	Knowledge Cutoff	Notes for Replication
ChatGPT	Bing-based browsing tool	GPT-5 or GPT-4o	Early-to-mid 2024	Used the standard ChatGPT web app, paid tier with browsing on. Used the default chat interface, not GPTs / agents.
Claude	Anthropic's web_search + web_fetch tools	Claude Sonnet 4.x	Late 2024 or May 2025	Standard Claude.ai web app.
Copilot	Bing-based browsing	Microsoft Prometheus stack built on GPT-4o or GPT-4 Turbo	Late 2023 to mid 2024	Standard Copilot Chat (consumer), Creative/Balanced mode default.
Gemini	Google Search tool	Gemini 2.5 Pro	Late 2023 to mid 2024	Used the Gemini consumer app. Engine audited as having did NOT utilize Google search tool or access live external databases across all nine trials.
Grok	xAI web search	Grok 4	Late 2023	Used SuperGrok on a paid subscription.
Perplexity	Perplexity's own crawler + web index	Sonar (or Sonar Large)	Late 2023 to mid 2024	Used the default Perplexity chat mode (not Pro Search / Copilot Search).

Replication Parameters

Period	May 8-9, 2026
Location	Mississauga, Ontario, Canada (engines that disclose regional inference flagged Canadian regional settings)
Sessions	Fresh browser session per trial. Three independent trials per engine, with the engine logged out and restarted between trials where possible.
Mode	Default consumer chat mode for each engine. No deep research, thinking, or agent modes engaged unless that was the engine default.
Prompt sequence	P1 to P2 to P3 to P4 to P5 in a single session, pasted verbatim, no follow-up questions between them.

Source: Buckler Dispersion Study. Controlled comparison of six consumer AI engines applied to Canadian-equity, equity-mutual-fund, and ETF top-five screening prompts. Data collected May 8-9, 2026 from Mississauga, Ontario.