

Tax Research Tested: How specialist AI compares with the leading AI labs on Australian tax

We tested six AI research systems on 91 Australian tax questions under current law. Each answer was graded against an expert reference answer.

Summary

Nylon scored highest of the six systems tested, ahead of Anthropic Claude Fable 5.1 and OpenAI GPT-6 Astra, and scored significantly higher on correctness than every other system.

We tested six AI research systems on 91 Australian tax questions drawn from the CPA Program's Taxation review questions, including only questions whose answers still hold under the law in force today. Every answer was graded against an expert reference answer by an independent judge model, using a five-part rubric that reflects how a reviewing tax practitioner checks draft advice.

Nylon scored a composite of **4.68 out of 5**, against 4.32 for Claude Fable 5.1 and 4.28 for GPT-6 Astra. Its lead over every other system is statistically significant on the composite and on correctness, faithfulness, helpfulness and citation grounding, and it scored highest on completeness.

On speed, Nylon's mean time to completion was 108 seconds, close to the other systems (98–126 seconds). Its slowest answers were also joint-fastest of any system, with 90% complete within 135 seconds, level with Gemini 3.1 Pro.

The full benchmark, including question fingerprints, the rubric, the judge model's instructions, every answer, every verdict and a script that recomputes every number, is available for independent review.

Why We Created This Benchmark

01

Why We Created This Benchmark

Tax advisers are adopting AI research tools quickly, but they have had no common yardstick for comparing them. General legal and reasoning benchmarks do not test what a practitioner actually needs, which is whether an answer applies the right section of the right Act, for the right period, to the facts in front of it.

To our knowledge, there is no established benchmark for AI tax research in Australia or New Zealand. So we created one, using an established set of Australian tax questions with expert reference answers, and graded every system the same way.

Our hypothesis

We anticipated that Nylon's specialist research harness, its mapping of legal data to the way practitioners find authority, and its purpose-built research tools would give it an advantage over the leading general-purpose AI systems. The benchmark was designed to test this against the same questions, tools and grading for every system.

Why firms choose Nylon

This benchmark tests only the task every system here can perform, which is researching and writing an answer to a single question. Nylon's other capabilities, which provide much of the value firms get from Nylon, were not relevant to or used in this benchmark. They include:

- **Non-public, government-origin data**, including scanned case law
- **An internal precedent system** that is updated frequently
- **An estimated certainty score** on every answer
- **Single-interface verification** of every source
- **Siloed engagement context** that keeps client data secure
- **A specialist tax notetaker** which spots tax problems in client calls

Method

02

Method

Questions

The questions are the 120 end-of-module review questions, with suggested answers, in CPA Program Australia Taxation, 6th edition (2023). They were extracted from the book's text layer, module by module, with each question transcribed together with its back-of-book suggested answer and checked against a fingerprint record of the extraction. No manual editing was applied. The textbook's answers apply the law for the 2022–23 income year, but tax law has moved on. Because the benchmark keeps only questions whose reference answer still holds under the law in force today, and grades each answer against that reference, we excluded 29 questions before any system's answers were scored:

- **21 questions** where the law has changed since 2022–23 (for example the replacement of the AAT by the Administrative Review Tribunal, the end of temporary full expensing, the Stage 3 rate changes and indexed thresholds), so the reference answer is no longer correct.
- **8 questions** where the reference answer is incorrect under the law it applies, confirmed against primary sources.

Each exclusion is listed with its reason in Appendix A. The remaining 91 questions cover all areas of tax, including tax administration and ethics, income and deductions, capital gains tax, individuals, entities (companies, partnerships and trusts), GST and fringe benefits tax.

Systems

System	How it was run
Nylon (October Snapshot)	A snapshot of Nylon's standard research service
OpenAI GPT-6 Astra	Provider API with web search and page reading
Anthropic Claude Fable 5.1	Provider API with web search and page reading
Anthropic Claude Opus 5.5	Provider API with web search and page reading
OpenAI GPT-6 Sol	Provider API with web search and page reading
Google Gemini 3.1 Pro	Provider API with web search and page reading

The systems were run between 28 September and 1 October 2026. The five comparison systems used the same web search and page-reading tools. OpenAI's API caps tool calls at 20 per question, so GPT-6 Astra and GPT-6 Sol were limited to 20. Gemini and Claude had no ceiling. GPT-6 Sol hit the ceiling on 52 of 91 questions.

Grading

Every answer was graded by a judge model, Kimi K3. It is made by Moonshot AI, which has no system in this test. The judge model received the question, the reference answer, the answer to be graded, and the text each system actually retrieved for every source it cited, so it could check each claim against the material the system relied on. It was not told which system wrote the answer, and graded each answer in a separate prompt.

The rubric, the Nylon Practitioner Rubric v1.0 (Appendix B), grades five dimensions the way a reviewing tax practitioner reviews draft advice. Each dimension is scored from 1 to 5. The percentage next to each dimension is its weight in the composite score, which is the weighted average of the five scores.

1. **Correctness (15%).** The right conclusions, rules and figures, under the law for the period the facts fix. A statutory rule should be stated from the legislation itself, with its operative words (not only from a guidance page, commentary or a ruling that restates the section), because the Act is the authority a practitioner relies on.
2. **Completeness (10%).** Every key point, condition, exception and qualification, with computations shown as workings and the leading judgment engaged where judicial authority decides the point.
3. **Faithfulness (30%).** The share of load-bearing claims that are anchored to a cited source whose retrieved text supports them, with the governing words set out so the reader need not take a rule on trust.
4. **Helpfulness (10%).** Form and explanation, including a direct answer for the period in issue, clear structure, computations laid out, the governing words set out, and a concrete next step.
5. **Citation grounding (35%).** Whether claims are anchored, at the claim, to relevant authority whose retrieved text verifies them, with statutory propositions grounded in the legislation.

Faithfulness and citation grounding carry the most weight because an answer a practitioner cannot verify cannot be relied on.

Speed

Speed is the time in seconds from request to the last word of the answer. Response times depend on provider load at the time of the run.

Results

03

Results



Figure 1: Composite score (Nylon Practitioner Rubric v1.0, judged by Kimi K3).

Nylon ranked first with a composite score of **4.68**. Claude Fable 5.1 (4.32) and GPT-6 Astra (4.28) came second and third with little between them, followed by Claude Opus 5.5 (4.07), GPT-6 Sol (3.96) and Gemini 3.1 Pro (2.97).

System	Correct.	Complete.	Helpful.	Faithful.	Grounding	Composite
Nylon	4.46	4.63	4.85	4.75	4.67	4.675
Claude Fable 5.1	4.20	4.57	4.40	4.33	4.27	4.321
GPT-6 Astra	4.08	4.46	3.81	4.32	4.41	4.277
Claude Opus 5.5	3.80	4.32	4.20	4.02	4.12	4.071
GPT-6 Sol	3.78	4.15	3.47	3.97	4.11	3.958
Gemini 3.1 Pro	3.24	3.98	2.91	2.51	2.97	2.965

Head to head

This table shows the difference between Nylon’s score and the average score of each other system. A positive score indicates that Nylon scored higher.

Nylon vs	Correct.	Complete.	Helpful.	Faithful.	Grounding	Composite
Claude Fable 5.1	+0.26	+0.05	+0.45	+0.42	+0.40	+0.35
GPT-6 Astra	+0.38	+0.16	+1.03	+0.43	+0.26	+0.40
Claude Opus 5.5	+0.66	+0.31	+0.65	+0.73	+0.55	+0.60
GPT-6 Sol	+0.68	+0.47	+1.37	+0.78	+0.56	+0.72
Gemini 3.1 Pro	+1.22	+0.65	+1.93	+2.24	+1.70	+1.71

Every difference is statistically significant at 95% confidence except completeness against Claude Fable 5.1 (+0.05). On correctness, Nylon had a statistically significant lead over the next two systems.

Speed

For practitioners, consistency in response time is more valuable than average speed, so a predictable wait can be planned around a client call or a review. The 90th percentile (the time by which nine in ten answers were complete) is therefore the most relevant speed measure.

System	Mean time (s)	90th percentile (s)
Nylon	108	135
Claude Fable 5.1	126	210
GPT-6 Astra	106	201
Claude Opus 5.5	101	164
GPT-6 Sol	98	139
Gemini 3.1 Pro	101	135

Nylon’s mean time to completion was close to the other systems, and its slowest answers were joint-fastest. 90% of Nylon’s answers were complete within 135 seconds, level with Gemini 3.1 Pro, against 201 seconds for Astra, 164 seconds for Opus and 210 seconds for Fable.

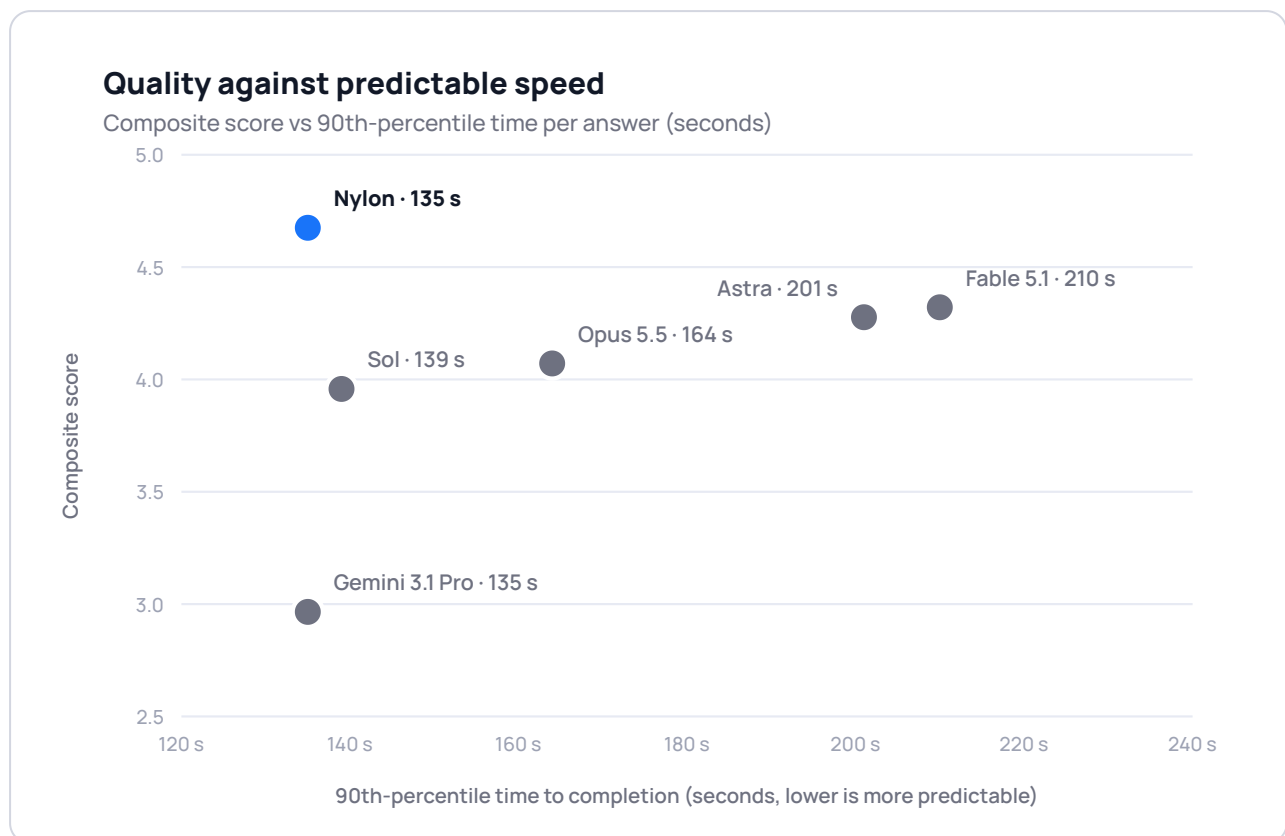


Figure 2: Nylon combines the highest score with the joint-fastest 90th-percentile time.

Discussion

04

Discussion

What the results show

62% of Nylon's answers scored a perfect 5 out of 5 on correctness, against 40% for Claude Fable 5.1 and 2–18% for the other systems, significantly more often than every other system. The specialist design paid off where a practitioner most needs it. Nylon's answers were the easiest to verify because its claims were anchored to authority whose text supports them, with statutory rules taken from the legislation itself.

The general-purpose systems more often relied on guidance pages or secondary summaries for statutory rules, or made claims their sources did not bear out.

A point-in-time test

These results are a snapshot. AI models change every week, and the systems tested here will change too. We see it as our responsibility to our customers to stay ahead of those changes, so Nylon will continue to improve each week.

Other specialist tools

Other specialist tools in Australia and New Zealand could not be included, as it would be against their terms of use. In our preliminary testing, we did not find any other specialist research AI that outperformed GPT-6 Astra or Claude Fable 5.1.

What comes next

Further to these results, we expect the strongest performance to come from combining the leading labs' agents with Nylon, through Nylon's MCP. We will publish results from this as a follow-up report.

The full benchmark is available for independent review. It includes question fingerprints, exclusions, the rubric, the judge model's exact instructions, every answer, every verdict with the judge model's reasons, and a script that recomputes every number in this paper. More detail in Appendix C.

Appendices

Appendix A: Excluded Questions

Both exclusion screens were applied to the questions and reference answers alone, before any system's answers were scored.

A.1 Law changed since 2022–23

21 questions.

Question	What changed
q1.13	AAT replaced by the Administrative Review Tribunal (Oct 2024); Small Taxation Claims Tribunal abolished
q1.14	Review and appeal now under the ART Act 2024
q1.16	Commonwealth penalty unit increased
q2.15	Cents-per-kilometre rate changed
q2.18	Temporary full expensing ended 30 June 2023
q2.22	Temporary full expensing ended; car limit changed
q2.23	TFE replaced by the \$20,000 instant asset write-off
q2.26	Temporary full expensing ended 30 June 2023
q4.2	Resident tax rates changed (Stage 3; 15% from 1 July 2026)
q4.5	Concessional contributions cap and SG rate changed
q4.6	Low-rate cap indexed; preservation age now 60
q4.8	Dependant offset amounts CPI-indexed

Question	What changed
q4.10	Dependant offset and spouse threshold CPI-indexed
q4.12	Private health insurance rebate tiers and percentages changed
q4.13	Rate on the \$18,201-\$45,000 band reduced
q4.14	Medicare levy surcharge thresholds indexed
q4.15	Tax rates, SG rate, cents-per-km rate and PHI rebate changed
q4.16	Resident tax rates changed (Stage 3; 15% from 1 July 2026)
q6.14	FBT benchmark interest rate changed
q6.18	Resident tax rates changed (Stage 3; 15% from 1 July 2026)
q6.20	FBT benchmark interest rate changed

A.2 Reference answer incorrect

8 questions where the reference answer is incorrect under the law it applies, confirmed against primary sources. The authority for the correct position is given for each.

Question	Defect in the reference answer	Authority
q1.18	Treats the full \$200/hour as assessable; s 26-35(4) ITAA 1997 makes the denied amount neither assessable nor exempt	ITAA 1997 s 26-35(1), (4)
q2.14	Uses the withdrawn 52c/hour home-office rate; PCG 2023/1 applies from 1 July 2022	PCG 2023/1
q2.20	Applies s 40-880(2A) to an existing business; s 40-880(2) applies	ITAA 1997 s 40-880(2), (2A)
q3.13	Says replacement shares keep pre-CGT status after a Subdiv 124-M rollover	ITAA 1997 s 104-10(5), Subdiv 124-M
q5.12	Treats a distribution bearing family trust distribution tax as assessable under s 97	ITAA 1936 Sch 2F; ITAA 1997 s 11-55
q6.2	Classifies land as goods; land is real property under s 195-1 GST Act	GST Act ss 9-10(2), 195-1
q6.7	Denies the associate an input tax credit available under Subdiv 72-B	GST Act ss 72-40, 72-45
q6.17	Aggregates associated benefits against the \$300 minor benefit threshold	FBTAA s 58P(1)(e), (f); TR 2007/12

Appendix B: Nylon Practitioner Rubric v1.0

The judge model received the following standard together with its general grading instructions (verification of claims against retrieved text; claim-level anchoring; primacy of the legislation for statutory propositions). Each dimension is scored 1–5; the levels below are given to the judge model verbatim.

Grade the answer as a reviewing partner would grade a draft advice. The strongest answers show their law and their working: they state each statutory rule from the legislation itself and set out its operative words in a short quotation, rather than relying on a revenue-authority page, commentary or a ruling that restates the section; they lay any computation out as a workings table a reviewer can follow step by step; where judicial authority governs a point, they engage the leading case and quote the passage of the judgment that decides it; they apply the law for the period the facts fix; they assert nothing the sources in front of the reader do not bear out; and they close with a specific next step – the authority to read next or the action to take. Correctness and Completeness are about substance: an answer earns nothing on either for presenting its conclusion in a separate conclusion or summary section, or for writing section references out in the prose – a rule identified through its citation counts as identified. Narration of the assistant's own research process ("I'll research...", "Let me search...", "I found...") adds nothing to what the answer establishes and counts against both. These are the marks of advice that can be relied on and checked, and the level descriptions below build them in. Read each dimension's levels as a ladder: an answer reaches a level only when it meets everything that level describes, and a trait a level names is required at that level and above. Helpfulness concerns form and explanation – how usable and well-explained the answer is to a professional – and is scored separately from whether its conclusions are right. Mark each dimension strictly: a 5 means nothing a reviewing partner would change on that dimension.

Correctness (15% of composite)

Level	Criteria
1	Reaches the wrong conclusion, or relies on a rule/section that contradicts the reference.
2	A material error; reliance on superseded law or on another period's rates or rules where that changes a figure or conclusion; statutory rules stated only through secondary material rather than the legislation; or a load-bearing claim that the captured sources do not support.
3	Right direction and broadly current, with each statutory rule stated from the legislation itself; every load-bearing claim borne out by the captured sources, though a minor claim may go beyond them; a minor imprecision in a rule, section, figure or effective date.
4	Right conclusions with the correct rules and sections for the period in issue, each statutory rule stated from the legislation itself and every claim borne out by the captured sources; small slips only; no narration of the assistant's own research process.
5	Right conclusions with the correct rules, sections and figures throughout; the law for the period the facts fix correctly applied; each load-bearing statutory rule stated from the legislation with its operative words quoted; nothing asserted that the captured sources do not bear out; no narration of the assistant's own research process.

Completeness (10% of composite)

Level	Criteria
1	Misses nearly all key points, conditions, exceptions and qualifications.
2	Covers only a small fraction of the key points, or major limbs of the reasoning are absent.
3	Core points covered, with the judicial authority on the area at least referred to; several conditions, exceptions or qualifications omitted.
4	Almost all key points present, with the leading judgment engaged and a passage of it quoted; only minor qualifications missing; no narration of the assistant's own research process.
5	Every key point, condition, exception and qualification in the reference is covered; every computation shown as a workings the reader can follow; the leading judgment engaged and quoted where it decides the point; no narration of the assistant's own research process.

Faithfulness (30% of composite)

Level	Criteria
1	Under 25% of load-bearing claims are anchored to a cited source whose captured text verifies them; assertions largely float free or rest on unverified references.
2	25-49% of load-bearing claims anchored and verified from captured text; largely ungrounded or reliant on the grader to construct the claim-to-source mapping.
3	50-74% of load-bearing claims anchored and verified, and the operative words of the governing legislation (or of the governing professional standard or code, where the question turns on one) set out verbatim so the reader need not take the rule on trust; one or more material claims unanchored or unverified.
4	75-94% of load-bearing claims anchored at the claim and verified from captured text, with the governing statutory words set out verbatim; at most a minor, low-importance claim unverified.
5	95%+ of load-bearing claims anchored at the claim to a cited source whose captured text verifies them, with the governing statutory words set out verbatim; a reviewer could audit the whole answer without leaving the page.

Helpfulness (10% of composite)

Level	Criteria
1	Does not usefully address the question.
2	Hard to use as a professional deliverable: off-point, consumer-grade, or lacking the working, the statutory text or the next step that level 3 requires.
3	Useful, and has the working elements of professional advice: the governing statutory words set out, any computation laid out as a workings table, and a concrete closing next step (the authority to read next or the action to take); but contains off-point material, is disorganised, or is mis-pitched.
4	Has those working elements and answers well; minor structure or tone issues; relevant caveats are not held against it.
5	Has those working elements; directly answers for the period in issue; tightly structured and correctly pitched to a professional adviser; includes the caveats a professional would want; no padding.

Citation grounding (35% of composite)

Level	Criteria
1	No relevant authority cited, or citations unanchored AND unverified: nothing in the answer can be audited from the material provided.
2	Few claims auditable: citations largely unanchored, unverified (teasers, tables of contents, truncated or uncaptured text), secondary-only for statutory propositions, citing no legislation at all, or unsupportive of the claims citing them.
3	Key claims cite relevant authority, but grounding is set-level rather than claim-anchored, or material claims rest on unverified captures or on secondary guidance where the statute itself was required. (An otherwise-strong bibliography without claim-level anchoring belongs here, capped at 3.)
4	Claims densely anchored at the claim to relevant authority, mostly verified from captured primary text; at most a small minority unverified or secondary-only, and at most trivial irrelevant citations.
5	Every material claim anchored at the claim to relevant authority whose captured text verifies it, with statutory propositions grounded in the captured words of the legislation itself; dense multi-source verified grounding with no irrelevant or unsupportive citations.

Appendix C: Reproducing the Results

The benchmark folder contains, for each of the 91 questions, its id and topic with a SHA-256 fingerprint of the exact question text and suggested answer used, and that text itself in encrypted form (it is the publishers' copyright; anyone holding the book can confirm the items match); the exclusion list with the reason for each; the rubric and the rationale for each of its elements; the judge model's system prompt, output schema and call settings; every system's answers (encrypted) with timings; every verdict with the judge model's reasons; the message the judge model received for each answer; per-question scores; and a register of every design decision with its justification, known cost and source. Running `scripts/score.py` recomputes every table in this paper from those files, `scripts/checks.py` recomputes the supporting checks, and `scripts/answers_crypto.py` decrypts the questions and answers with the key. Nylon's cost per answer is not included.

Repository

The benchmark folder is published on GitHub at github.com/LawCyborg/AU-Tax-Benchmark-Nylon.

Due to the copyrighted nature of the benchmark questions, we have encrypted the questions and answers, as the answers may contain copyrighted material. The systems' answers are also encrypted with AES-256. Each decrypted question and reference answer matches its published fingerprint. However, we have detailed how to recreate the test in the folder's README, with the appropriate judging criteria. If you would like to request the key to the answers, please email support@usenylon.com with your agreement not to use the data to compete with CPA Australia. We have reached out to CPA Australia for permission to release the questions to the public domain. We will decrypt both questions and answers if they grant us permission.

The sources each system cited, every verdict with the judge model's reasons, and every score remain in plain text.

Contact Us:

Research Team

benchmark@usenylon.com



The reasoning engine of the trusted adviser